Investigation into the role of sequence-driven-features and amino acid indices for the prediction of structural classes of proteins

By

Mr. Sundeep Singh Nanuwa

Submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy of

De Montfort University

April 2013

Abstract

The work undertaken within this thesis is towards the development of a representative set of sequence driven features for the prediction of structural classes of proteins. Proteins are biological molecules that make living things function, to determine the function of a protein the structure must be known because the structure dictates its physical capabilities. A protein is generally classified into one of the four main structural classes, namely all-α, all-β, α + β or α / β, which are based on the arrangements and gross content of the secondary structure elements. Current methods manually assign the structural classes to the protein by manual inspection, which is a slow process. In order to address the problem, this thesis is concerned with the development of automated prediction of structural classes of proteins and extraction of a small but robust set of sequence driven features by using the amino acid indices. The first main study undertook a comprehensive analysis of the largest collection of sequence driven features, which includes an existing set of 1479 descriptor values grouped by ten different feature groups. The results show that composition based feature groups are the most representative towards the four main structural classes, achieving a predictive accuracy of 63.87%. This finding led to the second main study, development of the generalised amino acid composition method (GAAC), where amino acid index values are used to weigh corresponding amino acids. GAAC method results in a higher accuracy of 68.02%. The third study was to refine the amino acid indices database, which resulted in the highest accuracy of 75.52%. The main contributions from this thesis are the development of four computationally extracted sequence driven feature-sets based on the underused amino acid indices. Two of these methods, GAAC and the hybrid method have shown improvement over the usage of traditional sequence driven features in the context of smaller and refined feature sizes and classification accuracy. The development of six non-redundant novel sets of the amino acid indices dataset, of which each are more representative than the original database. Finally, the construction of two large 25% and 40% homology datasets consisting over 5000 and 7000 protein samples, respectively. A public webserver has been developed located at http://www.generalised-protein-sequence-features.com, which allows biologists and bioinformaticians to extract GAAC sequence driven features from any inputted protein sequence.

Keywords: protein structural classes, sequence driven features, amino acid indices, test procedures, largest protein structural class datasets, generalised amino acid composition

Publications

Nanuwa, S. S. and H. Seker (2008). Investigation into the role of sequence-driven-features for prediction of protein structural classes. 8th IEEE International Conference on Bioinformatics and BioEngineering, BIBE 2008, Athens Greece. Volume 1. Pages 583-586

Nanuwa, S. S., A. Dziurla and H. Seker. (2009). Weighted amino acid composition based on amino acid indices for prediction of protein structural classes. Final Program and Abstract Book - 9th International Conference on Information Technology and Applications in Biomedicine, ITAB 2009, Larnaca Cyprus. Volume 1. Pages 327-332

Nanuwa, S. S. and H. Seker (2011). Prediction of a protein's structural class using amino acid indices. The International Congress on Bioinformatics and Biomics, 2009 Izmir Turkey

This thesis is dedicated to my wife Sharan, dad Amarjit, mum Harjinder, Brother Jamie & the Nanuwa family.

Acknowledgements

I would like to offer my sincerest gratitude to my supervisor, Dr Huseyin Seker, who has supported me throughout my Masters and PhD. The work undertaken is through his encouragement, which gave me the determination to complete the thesis. One simply could not wish for a better supervisor. I would like to also offer my sincerest gratitude to my current place of work The Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory (JDRF/WT DIL), centred at the Cambridge Institute for Medical Research at University of Cambridge, allowing me to have the time needed to complete the thesis.

Table of Contents

List of Tables

List of Figures

Acronyms

Acronym	Acronym meaning
AAC	Amino Acid Composition
AAI	Amino Acid Indices
DNA	Deoxyribonucleic Acid
GAAC	Generalised Amino Acid Composition
KNN	K Nearest Neighbour
MKNN	Multiple K Nearest Neighbour
PCA	Principal Component Analysis
PDB	Protein Data Bank
PSC	Protein Structural Class(es)
RNA	Ribonucleic acid
SDF	Sequence Driven Feature(s)
SRM	Sequence Representation Matrix
SCOP	Structural Classification of Proteins
SVM	Support Vector Machine
3D	Three Dimensional
tDNA	Transfer DNA

Mathematical Symbols

Symbol	Definition
λ	Lambda
D	Euclidean distance
q	Query Test Protein Sample
p	Training Protein
k	*K* Nearest Neighbours
F(r)	Fraction (*F*) of each Amino Acid Residue *r*
r	Amino Acid of type *r*
Nr	Number of Amino Acid Residues of type *r*
F(r,s)	Fraction (*F*) of Amino Acid type r and s
Nrs	Number of Dipeptides of Amino Acid Type *r* and *s*
d	Autocorrelation Lag
Pi	Amino Acid Index Property at Position *i*
N	Number of Amino Acid Residues in a given Protein Sequence
w	Weighting Factor
$seqdi, i + d$	distance between the amino acids at positions *i* and *i+seqd*
R	Amino Acid Residue
x	Feature *x*
y	Feature *y*
$Prob$	Joint probabilistic distribution of feature x and feature *y*
Ω	Dataset of all the features
$\|S\|$	Number of features from the dataset of features
S	Selected subset of the features
$c= c_1, c_2, c_3, c_4$	Class labels of the four main structural classes of proteins
MI	Mutual Information
V_{MI}	Maximum Relevance Value

Chapter 1 - Introduction

1.1 Background

Proteins are the biological machines that make living things function and are found in every cell and tissue of our body and are made out of 20 different types of amino acids. These amino acids are chemical compounds that are made up from carbon, hydrogen, oxygen and nitrogen elements, which when combined in many different arrangements form various types of proteins that organisms need to function correctly. The different arrangements are made up of varying lengths of number of amino acids to form an amino acid sequence, which dictates what the proteins structure and function is – this is called the proteins primary structure. There are many different types of proteins, which altogether play a pivotal role in the function of an organism, examples, the protein collagen is vital for the strength, elasticity and composition of our hair and skin, insulin protein is crucial to controlling blood sugar levels and the protein alcohol dehydrogenase detoxifies alcohol (Purves 2001). Most medical drugs target known proteins to modify its structure and function of the protein in aid of health benefits, the drugs are dependent on an active binding site(s) on the protein which is dependent on physical shape of the protein known as the tertiary structure (Chou and Zhang 1995; Moll and Kavraki 2008). There are many properties to a protein such as its structure, mass, melting temperature, amino acid composition and amino acid sequence order, to name but a few, however, the main property of a protein is its function as this defines what the protein does. To identify the function of a protein, the structure of it must be known beforehand. The identification of proteins tertiary structure is a complex problem to solve in bioinformatics because there is only one possible physical shape for the protein to fold into from a vast number of possibilities. The number of possibilities is linked to the size of the proteins primary structure as the arrangement of amino acids when folded can form many different physical shapes but the most suitable fold is the one that takes minimum amount of energy to fold. The search space to find the tertiary structure of the protein is so vast that using computational resources to through each possibility will ever end and such resource are limited. To help reduce the search space as much prior information or properties are required to computationally model proteins, and one such piece of property is the structural class of the

protein, which if known can help deduce the overall folding pattern of a protein. The structural class of a protein is one of the most important property for characterising the overall folding type of a protein and plays an important role in protein function analysis and in drug design (Chou and Zhang 1995; Zhou and Assa-Munt 2001; Chou 2005).

Current lab based methods to determine the structure and function of a protein is done through manual methods such as x-ray diffraction or nuclear magnetic resonance (NMR), which are accurate in the field but are a slow process and cam be tedious for variety of proteins that are hard to capture through these methods (Dubchak, Muchink et al. 1995). The cost and time required acquiring protein structure and function through lab-based methods is very high, thus, computational methods are needed to supplement lab-based efforts. One of the reasons lab based methods are becoming slower is that the protein primary sequence information grows significantly faster than 100% experimentally defining a protein complete structure and function. The gap between verified tertiary protein structures and unverified protein primary sequence as of February 2013 is 88,170 experimentally verified tertiary protein structures and 27,834,581 non-experimentally verified protein amino acid sequences. Protein's primary sequences are being discovered faster than ever before and are being deposited into proteomic databases waiting to be verified upon, within unverified data, there is new knowledge to be extracted, which could have health benefits in the widest range of areas. Therefore, it is essential to develop computational methods to help speed up the process, and one such progress is to develop tools to predict the structural classes of proteins based on their primary protein structure.

There have been many attempts to use the primary protein structure to derive sequence driven features to predict structural classes of the protein, (Eisenhaber, Frömmel et al. 1996; Tomii and Kanehisa 1996; Bahar, Atilgan et al. 1997; Chou and Maggiora 1998; Wang and Yuan 2000; Cai, Liu et al. 2001; Chou and Cai 2004; Du, Jiang et al. 2006; Kyoung Kim, Bang et al. 2006; Xiao, Shao et al. 2006; Huang, Kawashima et al. 2007; Kurgan, Stach et al. 2007; Ong, Lin et al. 2007; Chen, Chen et al. 2008; Li, Zhou et al. 2008; Gu and Chen 2009; Yang, Peng et al. 2010; Saha, Maulik et al. 2011; Ding, Zhang et al. 2012). Among these studies, there are varying ranges of results obtained from many different types of datasets of varying size and homology coupled with using many different types of sequence driven features of varying sizes and complexity. What has not been identified in the field from all varying studies is a consensus on what is best type of sequence driven feature(s) suited for the prediction of

structural classes of proteins. This will be done through an investigation of into the largest set of sequence driven feature and amino acid indices (AAI) through a consistent methodology of using multiple high quality datasets and three different test procedures. The aim is to better understand the different types of sequence-driven feature groups and amino acid indices which are better suited to the prediction of structural classes of proteins and then to analyse and evaluate the effects of various factors such as dataset size, dataset homology and test procedures methods. The investigation will lead into to the development of novel several sequences driven features that are suited for prediction of structural classes and other application areas.

1.2 Prediction of structural classes of proteins

The structural class of a protein is a property used to characterise the overall folding type of a protein. The concept of protein structural class was developed over four decades ago based on a visual inspection of polypeptide chain topologies in a dataset of only 31 globular proteins (Levitt and Chothia 1976). During the past three decades, there have been many different methods proposed to predict structural classes of protein from using its amino acid sequence, these methods are built upon a variety of different models and each is tailored to serve the purpose for which it was built. Their methods developed mainly differ based on a single dataset, single evaluation method and a complex feature representation (Kurgan and Homaeian 2006). The four major factors that should be considered for prediction are (1) the sequence representation, (2) selection of the classification algorithm, (3) selection of the test procedure and (4) the selection of dataset which encompasses the size and homology levels of the dataset (Kurgan and Homaeian 2006). Each one these factor affects the classification, where and many studies only look into a single combination of these factors this thesis will investigate multiples of them.

Prediction accuracies in the literature have reached up to 90-100%, this may seem like it is a good result however the results are based on using high homology datasets where similarity between protein sequence are higher than 50% and the dataset size are between 200 and 600 protein these and cannot be considered reliable. However, low-homology datasets where the sequence homology are less than 40% the reported accuracies range between 50-70% - using good quality datasets, such as the widely used 25PDB and 1189 where sequence similarities are between 25% and 40% and dataset sizes are 1668 and 1089 respectively. Selection of appropriate and un-biased datasets is crucial to any classification method, as the datasets

must not over represent a single factor, which could lead to false predictive accuracy. In the literature the two most common factors that leads to unreliable results is (1) homology of dataset i.e. the similarity between data and (2) the dataset size. It was found that different datasets and test procedures produce different sets of results and that a consensus set of results should be obtained through each method.

1.3 Sequence-driven-features & amino acid indices

Often the results presented in the literature are achieved using complex sequence-driven feature representations that run into hundreds and even thousands of descriptors. The current literature has not fully investigated the full set of available sequence-driven-features for the prediction of structural classes in one study. The analysis of the largest set of sequence-driven-features will shed light on the features that are suited towards the prediction of structural classes of protein. The study also shed light on composition based feature groups which was overall better at predicting structural classes of proteins than any other type i.e. autocorrelation or sequence-order. The widely used composition feature which represents a protein as a twenty-dimensional vector corresponding to the frequencies of the twenty amino acids in a given protein amino acid sequence is the amino acid composition (AAC). However, AAC feature group assigns each amino acid type a single weight of one and ignores the importance of individual amino acids weights (or properties) that are available. It was found that studies such as (Chen, Tian et al. 2006; Xiao, Shao et al. 2006; Chou 2011) utilised such weights within its sequence driven feature set to derive its descriptor value which has been shown to be beneficial for their prediction tasks, however they utilises a very limited number of amino acid indices and does not take into consideration the generalisation abilities by utilising all of them. Amino acid indices are numerical values representing various physicochemical and biochemical properties of amino acids and pairs of amino acid. This thesis has (1) expanded the amino acid indices dataset to the largest collection of 611 non-redundant amino acid indices - the previous dataset contributed by Kawashima et al (Kawashima, Pokarowski et al. 2008) contained 544 which had not been updated since 2008. To overcome the limitation of traditional AAC, a new method was developed named generalised amino acid composition (GAAC) which replaces an artificial weight of 1 with natural biologically related weight in the form of amino acid indices. The results obtained have shown to better compared to traditional AAC and pseudo amino acid composition (PseAAC) feature groups. The PseAAC feature groups incudes a sequence-order descriptor, which supposedly captures the importance of the order of the amino acid residues in a protein

sequence, results show that traditional amino acid composition and GAAC performs better than PseAAC as a whole – in the context of structural classes of proteins.

One of the main investigation areas in this thesis involved the analysis of these amino acid indices and developing several novel sequence driven feature sets based on extracting features from the use of amino acid indices. The amino acid indices dataset was refined by clustering similar amino acid indices together and then deriving a summarised amino acid index by using principal component analysis (PCA). These new amino acid indices are computationally generated which produced better accuracies over the clustered amino indices it replaces and result one of the highest accuracy obtained for the low homology dataset 25PDB.

1.4 Organisation of PhD thesis

Having given a brief background, aims, objectives and the PhD research study in this first chapter, the thesis is organised in detail, the concepts, methods, materials experiments and achievements as follows.

- Chapter 2 introduces the broad areas surrounding the prediction of structural classes of protein, which introduces the biology behind proteins, current bioinformatics approaches to structural class prediction of proteins, current types of datasets used in the field and a review of the current sequence driven features including amino acid indices.
- Chapter 3 describes the detailed foundation and mathematics behind the materials and methods selected for the investigations. Construction of the datasets are discussed, the implementation of the classification algorithm multiple-k-nearest-neighbour, principal component analysis and hierarchical clustering are looked into details in the context of bioinformatics approaches.
- Chapter 4 undertakes the largest comprehensive analysis using the largest set of sequence-driven-features. It looks at which types of sequence driven features are best suited for the prediction of structural classes of proteins.
- Chapter 5 takes forward the findings presented in chapter 4 and extends the use of amino acid composition by presenting four novel sequences driven features utilising amino acid indices to develop four feature extraction methods to create novel computationally generated indices. It also presents a comprehensive analysis over the largest set of amino acid

indices with the aim to identify a candidate set of indices that are representative towards the prediction of structural classes of protein.

- Chapter 6 presents the feature selection work, which investigates the selection of a subset sequence driven features that better represents the original feature space using two widely used feature selection methods.
- Chapter 7 presents a discussion from the main body of work presented in chapter 4, 5 and 6, future work towards and a summary of the contributions of this thesis.

Chapter 2 - Literature Review

2.1 Introduction

Chapter 2 provides the biology behind structural classes of proteins and introduces the bioinformatics approaches to the prediction of structural classes of proteins. It contains backgrounds reading cells, deoxyribonucleic acid (DNA), protein sequence driven features, proteins, amino acid indices and structural classes of proteins datasets, which put together are the materials used for the classification of structural classes of proteins. This chapter provides the literature review of the progress made in the field and that factors that should be considered when addressing the area.

2.2 Cells, DNA, Proteins

Cells are the structural components of living organisms, in which billions of cells exist to serve varying biological functions such as:-

- Red blood cells – which pick up oxygen from the lungs and transports it around to other cells in the body.
- Nerve cells – which lay end to end and are the wiring of the body that send messages from one part to another.
- Muscle cells – which make up fibres of an organism's tissues.

The blueprint of life is encoded within the DNA (Feitelson and Treinin 2002) and inside every cell is a nucleus where this DNA is contained (Purves 2001). DNA is made of four chemical bases and contains the entire information needed to shape and define an organism. The four chemical bases that make up DNA are known as nucleotides adenine (A), cytosine (C), guanine (G) and thymine (T). The specific arrangement of an organism's DNA encodes the information for building an organism's genes. Genes encodes a unique protein that performs a specialised function in the cell such as transporting red blood cells to building an organism's tissues, the human genome contains more than 25,000 genes (Purves 2001). Genes are encoded by stretches of DNA sequences called exons, an exon is a sequence of DNA that is expressed (known as transcription) into ribonucleic acid (RNA), which is the portion of the gene DNA that dictates amino acids, which in turn encodes proteins. Cells use a two-step process called

transcription and translation to read each gene from the DNA and produces amino acids that make up the translated protein. The rules for translating a gene into a protein are set in the universal genetic code as shown in Table 2-1 (A. Brazma 2002).

The universal genetic code converts DNA into amino acids which are the building blocks of proteins (Gromiha and Selvaraj 1998). There are twenty naturally occurring amino acids that can be found in an organism that, when combined into a specific arrangement, can form an amino acid protein sequence of varying lengths that encodes a protein's various properties such as its structure and functionality (Chou 2004), this is called the primary structure of the protein. These twenty amino acids are listed in Table 2-2 along with their 3-letter and 1-letter abbreviation codes.

A protein is a biological molecule that carries out a specific function in the organism (Purves 2001) and are mainly classified into three different structural phases, which defines their properties and functionality (Rost 1998). They are (1) primary structure that is the sequence of amino acids, (2) secondary structure that forms the structural elements the protein is physically made up of and (3) tertiary structure that is regarded as the physical 3-demensional (3D) structure of a protein. Proteins give the organism the ability to carry out majority of the functions it needs such as carrying oxygen around the body (haemoglobin protein); lower the freezing point of water (antifreeze protein), detoxifying alcohol in the body (alcohol Dehydrogenase protein) and structural support (collagen protein), to name but a few.

The two steps that involve converting genes (DNA) into proteins are (1) transcription and (2) translation, together known as gene expression. Transcription transfers DNA into ribonucleic acid (RNA) in the cell's nucleus, RNA contains the information and processes needed to make the actual physical proteins from DNA. The transfer of information is called messenger RNA (mRNA) as it transfers the encoded messages from the DNA to the area of a cell called cytoplasm where translation occurs (Purves 2001; Feitelson and Treinin 2002). The difference between RNA and DNA is that the chemical base thymine (T) is replaced with (U) uracil, the main job of RNA is to transfer the genetic code needed for the creation of proteins. Translation reads the mRNA sequence of three bases at a time, which is called a codon. An example of a codon is shown in Table 2-1 as one of the bold highlighted set of three letters.

A codon encodes a particular amino acid as listed in Table 2-2, once the codons are read and converted into amino acids transfer DNA (tDNA) puts together each amino acid to form the

primary structure of the protein (Eidhammer I 2005). The transcription and translation making proteins from DNA process is illustrated in Figure 2-1. An example of how to read the genetic code is to read a DNA sequence from left to right in Table 2-1. E.g., a DNA sequence of ATGAGC split into codons as ATG (codon 1) and AGC (codon 2). Follow the table from left to right and match each letter in the codon, this example is highlighted in bold and letters underlined to match the first codon and the italicised *Met [M]* is what matches the three-letter codon to the amino acid it encodes. An illustration of how DNA is converted to amino acids and then into a protein, the process of transcription and translation is shown in Figure 2-2.

Table 2-1 The DNA and RNA Genetic Code that constructs amino acids. (T and U are used in DNA and RNA, respectively) (Purves 2001)

Chemical bases	First letter in codon	Chemical bases								Third letter in codon	Chemical bases
		Second letter in codon									
			T or U		C		A		G		
		T or U	**TTT**	Phe [F]	**TCT**	Ser [S]	**TAT**	Tyr [Y]	**TGT**	Cys [C]	T or U
			TTC	Phe [F]	**TCC**	Ser [S]	**TAC**	Tyr [Y]	**TGC**	Cys [C]	C
			TTA	Leu [L]	**TCA**	Ser [S]	**TAA**	STOP	**TGA**	STOP	A
			TTG	Leu [L]	**TCG**	Ser [S]	**TAG**	STOP	**TGG**	Trp [W]	G
		C	**CTT**	Leu [L]	**CCT**	Pro [P]	**CAT**	His [H]	**CGT**	Arg [R]	T or U
			CTC	Leu [L]	**CCC**	Pro [P]	**CAC**	His [H]	**CGC**	Arg [R]	C
			CTA	Leu [L]	**CCA**	Pro [P]	**CAA**	Gln [Q]	**CGA**	Arg [R]	A
			CTG	Leu [L]	**CCG**	Pro [P]	**CAG**	Gln [Q]	**CGG**	Arg [R]	G
		A	**ATT**	Ile [I]	**ACT**	Thr [T]	**AAT**	Asn [N]	**AGT**	Ser [S]	T or U
			ATC	Ile [I]	**ACC**	Thr [T]	**AAC**	Asn [N]	**AGC**	Ser [S]	C
			ATA	Ile [I]	**ACA**	Thr [T]	**AAA**	Lys [K]	**AGA**	Arg [R]	A
			ATG	*Met [M]*	**ACG**	Thr [T]	**AAG**	Lys [K]	**AGG**	Arg [R]	**G**
		G	**GTT**	Val [V]	**GCT**	Ala [A]	**GAT**	Asp [D]	**GGT**	Gly [G]	T or U
			GTC	Val [V	**GCC**	Ala [A]	**GAC**	Asp [D]	**GGC**	Gly [G]	C
			GTA	Val [V	**GCA**	Ala [A]	**GAA**	Glu [E]	**GGA**	Gly [G]	A
			GTG	Val [V	**GCG**	Ala [A]	**GAG**	Glu [E]	**GGG**	Gly [G]	G

Table 2-2 the twenty amino acids and their abbreviations

Amino Acid	3-Letter code	1-Letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y

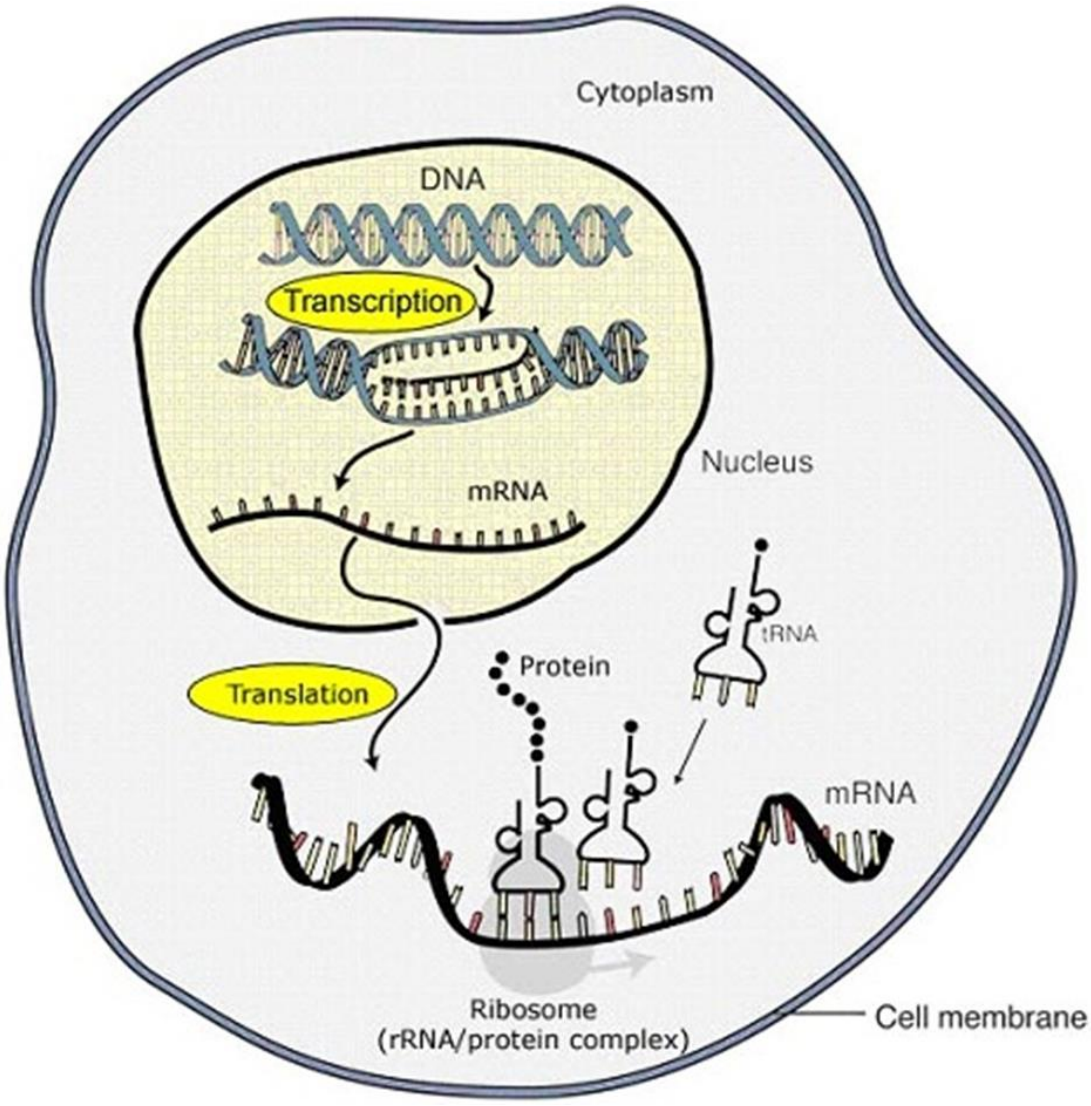


Figure 2-1 Transcription and Translation: Making proteins from DNA (Purves 2001)

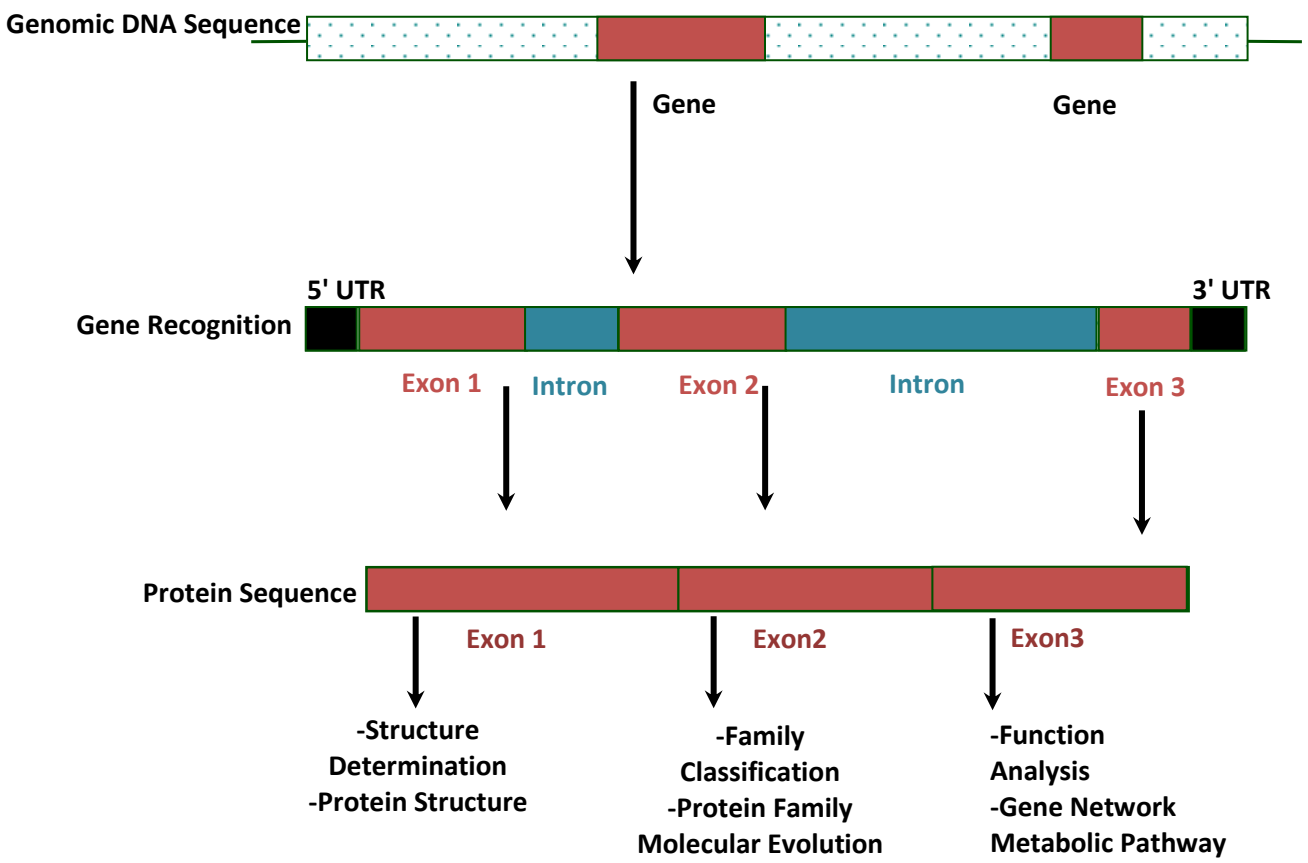


Figure 2-2 Illustration of how genes are obtained from DNA sequence

Figure 2-2 shows a stretch of DNA that is split between protein coding sequence called exons (red blocks) and non-protein coding parts called introns (Green blocks). The exons are joined together to form a gene. The introns are described as junk DNA (Purves 2001) and is largely ignored/removed. The three exons remaining (exon 1, exon 2 and exon 3) join to form a stretch of DNA that encodes a gene and which then in turn encodes for proteins.

Figure 2-3 is an example DNA sequence and belongs to the human alcohol dehydrogenase protein that detoxifies alcohol from the body. DNA to amino acid conversion of the sequence is shown in Table 2-3. The DNA sequence is read in codons, encoding particular amino acid as listed in Table 2-1. The first codon of this DNA sequence is **ATG**, which ncodes for the amino acid **Met [M]** methionine amino acid as seen in Table 2-3 its 3-letter and 1-letter abbreviations are Met and M, respectively. Similarly, this process is followed to form a protein sequence.

>DNA|AAB59496|AAB59496.1 Homo sapiens (human) alcohol dehydrogenase beta-1

ATGAGCACAGCAGGAAAAGTAATCAAATGCAAAGCAGCTGTGCTATGGGAGGTAAAGAAACCC
TTTTCCATTGAGGATGTGGAGGTTGCACCTCCTAAGGCTTATGAAGTTCGCATTAAGATGGTGGC
TGTAGGAATCTGTCGCAGATGACCACGTGGTTAGTGGCAACCTGGTGACCCCCCTTCCTGTGATT
TTAGGCCATGAGGCAGCCGGCATGTGGAGAGTGTTGGAGAAGGGGTGACTACAGTCAAACCAG
GTGATAAAGTCATCCCGCTCTTTACTCCTCAGTGTGGAAAATGCAGAGTTTGTAAAAACCCGGAG
AGCAACTACTGCTTGAAAAATGATCTAGGCAATCCTCGGGGGACCCTGCAGGATGGCACCAGGA
GGTTCACCTGCAGGGGGAAGCCCATTCACCACTTCCTTGGCACCAGCACTTCTCCCAGTACACGG
TGGTGGATGAGAATGCAGTGGCCAAAATTGATGCAGCCTCGCCCCTGGAGAAAGTCTGCCTCAT
TGGCTGTGGATTCTCGACTGGTTATGGGTCTGCAGTTAACGTTGCCAAGGTCACCCCAGGCTCTA
CCTGTGCTGTGTTTGGCCTGGGAGGGGTCGGCCTATCTGCTGTTATGGGCTGTAAAGCAGCTGG
AGCAGCCAGAATCATTGCGGTGGACATCAACAAGGACAAATTTGCAAAGGCCAAAGAGTTGGGT
GCCACTGAATGCATCAACCCTCAAGACTACAAGAAACCCATCCAGGAAGTGCTAAAGGAAATGA
CTGATGGAGGTGTGGATTTTTCGTTTGAAGTCATCGGTCGGCTTGACACCATGATGGCTTCCCTG
TTATGTTGTCATGAGGCATGTGGCACAAGCGTCATCGTAGGGGTACCTCCTGCTTCCCAGAACCT
CTCAATAAACCCTATGCTGCTACTGACTGGACGCACCTGGAAGGGGGCTGTTTATGGTGGCTTTA
AGAGTAAAGAAGGTATCCCAAAACTTGTGGCTGATTTTATGGCTAAGAAGTTTTCACTGGATGCG
TTAATAACCCATGTTTTACCTTTTGAAAAAATAAATGAAGGATTTGACCTGCTTCACTCTGGGAAA
AGTATCCGTACCGTCCTGACGTTTT

Figure 2-3 DNA sequence of human alcohol dehydrogenase protein taken from PDB (Davis, Bosron et al. 1996)

Table 2-3 DNA to amino acid for DNA sequence of human alcohol dehydrogenase given in Figure 2-3

Residue No.	DNA	Amino Acid	Residue No.	DNA	Amino Acid	Residue No.	DNA	Amino Acid	Residue No.	DNA	Amino Acid
1	AGT	S	48	CGC	T	95	CCT	P	142	GGC	G
2	AGC	T	49	AGA	D	96	CAG	Q	143	ACC	T
3	ACA	A	50	TGA	D	97	TGT	C	144	AGC	S
4	GCA	G	51	CCA	H	98	GGA	G	145	ACT	T
5	GGA	K	52	CGT	V	99	AAA	K	146	TCT	F
6	AAA	V	53	GGT	V	100	TGC	C	147	CCC	S
7	GTA	I	54	TAG	S	101	AGA	R	148	AGT	Q
8	ATC	K	55	TGG	G	102	GTT	V	149	ACA	Y
9	AAA	C	56	CAA	N	103	TGT	C	150	CGG	T
10	TGC	K	57	CCT	L	104	AAA	K	151	TGG	V
11	AAA	A	58	GGT	V	105	AAC	N	152	TGG	V
12	GCA	A	59	GAC	T	106	CCG	P	153	ATG	D
13	GCT	V	60	CCC	P	107	GAG	E	154	AGA	E
14	GTG	L	61	CCT	L	108	AGC	S	155	ATG	N
15	CTA	W	62	TCC	P	109	AAC	N	156	CAG	A
16	TGG	E	63	TGT	V	110	TAC	Y	157	TGG	V
17	GAG	V	64	GAT	I	111	TGC	C	158	CCA	A
18	GTA	K	65	TTT	L	112	TTG	L	159	AAA	K
19	AAG	K	66	AGG	G	113	AAA	K	160	TTG	I
20	AAA	P	67	CCA	H	114	AAT	N	161	ATG	D
21	CCC	F	68	TGA	E	115	GAT	D	162	CAG	A
22	TTT	S	69	GGC	A	116	CTA	L	163	CCT	A
23	TCC	I	70	AGC	A	117	GGC	G	164	CGC	S
24	ATT	E	71	CGG	G	118	AAT	N	165	CCC	P
25	GAG	D	72	CAT	I	119	CCT	P	166	TGG	L
26	GAT	V	73	GTG	V	120	CGG	R	167	AGA	E
27	GTG	E	74	GAG	E	121	GGG	G	168	AAG	K
28	GAG	V	75	AGT	S	122	ACC	T	169	TCT	V
29	GTT	A	76	GTT	V	123	CTG	L	170	GCC	C
30	GCA	P	77	GGA	G	124	CAG	Q	171	TCA	L
31	CCT	P	78	GAA	E	125	GAT	D	172	TTG	I
32	CCT	K	79	GGG	G	126	GGC	G	173	GCT	G
33	AAG	A	80	GTG	V	127	ACC	T	174	GTG	C
34	GCT	Y	81	ACT	T	128	AGG	R	175	GAT	G
35	TAT	E	82	ACA	T	129	AGG	R	176	TCT	F
36	GAA	V	83	GTC	V	130	TTC	F	177	CGA	S
37	GTT	R	84	AAA	K	131	ACC	T	178	CTG	T
38	CGC	I	85	CCA	P	132	TGC	C	179	GTT	G
39	ATT	K	86	GGT	G	133	AGG	R	180	ATG	Y
40	AAG	M	87	GAT	D	134	GGG	G	181	GGT	G
41	ATG	V	88	AAA	K	135	AAG	K	182	CTG	S
42	GTG	A	89	GTC	V	136	CCC	P	183	CAG	A
43	GCT	V	90	ATC	I	137	ATT	I	184	TTA	V
44	GTA	G	91	CCG	P	138	CAC	H	185	ACG	N
45	GGA	I	92	CTC	L	139	CAC	H	186	TTG	V
46	ATC	C	93	TTT	F	140	TTC	F	187	CCA	A
47	TGT	R	94	ACT	T	141	CTT	L	188	AGG	K
189	TCA	V	239	AAT	E	289	GCG	S	339	AGT	K
190	CCC	T	240	GCA	C	290	TCA	V	340	TTT	F
191	CAG	P	241	TCA	I	291	TCG	I	341	CAC	S
192	GCT	G	242	ACC	N	292	TAG	V	342	TGG	L
193	CTA	S	243	CTC	P	293	GGG	G	343	ATG	D
194	CCT	T	244	AAG	Q	294	TAC	V	344	CGT	A
195	GTG	C	245	ACT	D	295	CTC	P	345	TAA	L
196	CTG	A	246	ACA	Y	296	CTG	P	346	TAA	I
197	TGT	V	247	AGA	K	297	CTT	A	347	CCC	T
198	TTG	F	248	AAC	K	298	CCC	S	348	ATG	H

(Continued Table 2-3)

199	GCC	G	249	CCA	P	299	AGA	Q	349	TTT	V
200	TGG	L	250	TCC	I	300	ACC	N	350	TAC	L
201	GAG	G	251	AGG	Q	301	TCT	L	351	CTT	P
202	GGG	G	252	AAG	E	302	CAA	S	352	TTG	F
203	TCG	V	253	TGC	V	303	TAA	I	353	AAA	E
204	GCC	G	254	TAA	L	304	ACC	N	354	AAA	K
205	TAT	L	255	AGG	K	305	CTA	P	355	TAA	I
206	CTG	S	256	AAA	E	306	TGC	M	356	ATG	N
207	CTG	A	257	TGA	M	307	TGC	L	357	AAG	E
208	TTA	V	258	CTG	T	308	TAC	L	358	GAT	G
209	TGG	M	259	ATG	D	309	TGA	L	359	TTG	F
210	GCT	G	260	GAG	G	310	CTG	T	360	ACC	D
211	GTA	C	261	GTG	G	311	GAC	G	361	TGC	L
212	AAG	K	262	TGG	V	312	GCA	R	362	TTC	L
213	CAG	A	263	ATT	D	313	CCT	T	363	ACT	H
214	CTG	A	264	TTT	F	314	GGA	W	364	CTG	S
215	GAG	G	265	CGT	S	315	AGG	K	365	GGA	G
216	CAG	A	266	TTG	F	316	GGG	G	366	AAA	K
217	CCA	A	267	AAG	E	317	CTG	A	367	GTA	S
218	GAA	R	268	TCA	V	318	TTT	V	368	TCC	I
219	TCA	I	269	TCG	I	319	ATG	Y	369	GTA	C
220	TTG	I	270	GTC	G	320	GTG	G	370	CCG	T
221	CGG	A	271	GGC	R	321	GCT	G	371	TCC	V
222	TGG	V	272	TTG	L	322	TTA	F	372	TGA	L
223	ACA	D	273	ACA	D	323	AGA	K	373	CGT	T
224	TCA	I	274	CCA	T	324	GTA	S	374	TTT	F
225	ACA	N	275	TGA	M	325	AAG	K			
226	AGG	K	276	TGG	M	326	AAG	E			
227	ACA	D	277	CTT	A	327	GTA	G			
228	AAT	K	278	CCC	S	328	TCC	I			
229	TTG	F	279	TGT	L	329	CAA	P			
230	CAA	A	280	TAT	L	330	AAC	K			
231	AGG	K	281	GTT	C	331	TTG	L			
232	CCA	A	282	GTC	C	332	TGG	V			
233	AAG	K	283	ATG	H	333	CTG	A			
234	AGT	E	284	AGG	E	334	ATT	D			
235	TGG	L	285	CAT	A	335	TTA	F			
236	GTG	G	286	GTG	C	336	TGG	M			
237	CCA	A	287	GCA	G	337	CTA	A			
238	CTG	T	288	CAA	T	338	AGA	K			

2.2.1 Primary Structure

The basic form of a protein is its primary structure; which is a linear chain of amino acids generated by tDNA to form protein's amino acid sequence (Purves 2001). Each amino acid present in a protein's amino acid sequence is referred to as a residue. A proteins amino acid sequence can range from as small as two or more residues to sequences with hundreds or thousands residues. Sequences of up to 10 residues are called peptides whereas larger sequences are regarded as proteins (Gorga 2008).

The primary structure shown in Figure 2-4 is an example of a proteins amino acid sequence of 374 residues long. The sequence represents the human alcohol dehydrogenase protein, which is present in the liver (Davis, Bosron et al. 1996). Identification code of the protein is 1HTB and is taken from the PDB (Berman 2007).

>1HTB:A|Homo sapiens (human) alcohol dehydrogenase

MSTAGKVIKCKAAVLWEVKKPFSIEDVEVAPPKAYEVRIKMVAVGICRTDD
HVVSGNLVTPLPVILGHEAAGIVESVGEGVTTVKPGDKVIPLFTPQCGKCRV
CKNPESNYCLKNDLGNPRGTLQDGTRRFTCRGKPIHHFLGTSTFSQYTVVD
ENAVAKIDAASPLEKVCLIGCGFSTGYGSAVNVAKVTPGSTCAVFGLGGVG
LSAVMGCKAAGAARIIAVDINKDKFAKAKELGATECINPQDYKKPIQEVLKE
MTDGGVDFSFEVIGRLDTMMASLLCCHEACGTSVIVGVPPASQNLSINPM
LLLTGRTWKGAVYGGFKSKEGIPKLVADFMAKKFSLDALITHVLPFEKINEG
FDLLHSGKSICTVLTF

Figure 2-4 Protein Sequence for 1HTB taken from the Protein Data Bank (Davis, Bosron et al. 1996)

The primary structure is one of the most abundant pieces of proteomic information available, in almost all cases a protein will have a primary structure that is an amino acid sequence associated to it. The main data resource where primary structure information is available is the PDB (Bernstein, Koetzle et al. 1977; Berman 2007), UniProtKB / Swiss-Prot (Consortium 2012) and UniProtKB / TrEMBL (Consortium 2012). The PDB contains as of Tuesday Jan 08, 2013 there are 87279 verified protein structures. UniProtKB / Swiss-Prot which is manually annotated and reviewed as of October 2012 contains 538259 sequence entries. UniProtKB /

TrEMBL which is automatically annotated and is not reviewed as of October 2012 contains 27122814 sequence entries.

2.2.2 Secondary structure

Within the long protein chains (primary sequence), there are regions in which the chains are organised into regular structures known as alpha helices (α-helices) and beta-pleated sheets (β-sheets). This forms the secondary structures in proteins (Levitt and Chothia 1976; Nishikawa and Ooi 1980; Deleage and Roux 1987). Proteins do not stay in a linear form (primary structure) but ultimately fold into a tertiary structure, which the primary sequence encodes (Chou and Zhang 1995). The tertiary structure is made up of secondary structure elements (Lin and Pan 2001) – which is discussed in the section 2.2.3.

The two main types of secondary structure elements are α-helices and β-sheets. Figure 2-5 is an alpha-helix protein chain, which looks like a loosely coiled spring. The "alpha" means that if you look down the length of the spring, the coiling structure happens in a clockwise direction as it goes away from you (Kumarevel, Gromiha et al. 2000). Figure 2-6 shows a beta-pleated sheet protein chain folded so the strands lie alongside each other. Many α-helix elements is called α-helices (Chou and Zhang 1995). Many β-strands make up a β-sheet (Chou and Zhang 1995).

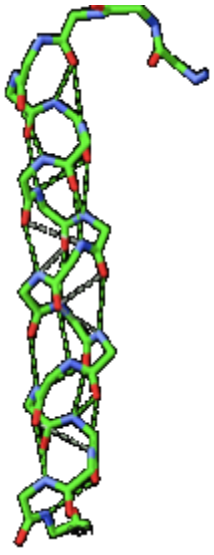


Figure 2-5 alpha-helix ribbon (Karadaghi 2012)

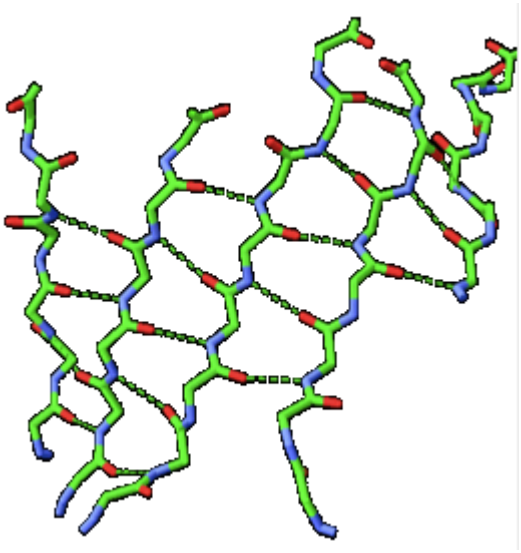


Figure 2-6 beta-sheet ribbon (Karadaghi 2012)

Anything that is not classified as α-helices or β-sheets is defined as coils, loops or turns. There is no universal agreement one the exact definition of these concepts (Ding, Zhang et al. 2009). They are defined by using the term loops (Eidhammer I 2005) and (Mizianty and Kurgan 2009; Yang, Peng et al. 2010) use the term coils. For the purpose of clear definition, the term coil will be used throughout this thesis. To form a complete protein the α-helices and/or β-sheets must be joined by using coils (Eidhammer I 2005). Every amino acid residue in proteins amino acid sequence can be classified into its secondary structure element, (h) helix, (e) strand and (c) coil; these are the contents of secondary structure (Yang, Peng et al. 2010). For example, the protein 1HTB has a domain with the amino acid sequence shown in Figure 2-4, the classified secondary structures is shown in Table 2-4. Domains are separate subunit functional and/or structural part of a protein. It is responsible for a particular function or interaction, contributing to the overall biological function of a protein. Domains may exist in a variety of biological contexts, where similar domains can be found in proteins with different functions (Bu, Feng et al. 1999).

Table 2-4 Classified secondary structure elements for the protein 1THB

SS	C	C	C	C	C	E	E	E	E	E	E	E	E	E	C	C	C	C	C	C
AA	S	T	A	G	K	V	I	K	C	K	A	A	V	L	W	E	V	K	K	P
SS	E	E	E	E	E	E	E	E	C	C	C	C	C	C	E	E	E	E	E	E
AA	F	S	I	E	D	V	E	V	A	P	P	K	A	Y	E	V	R	I	K	M
SS	E	E	E	E	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
AA	V	A	V	G	I	C	R	T	D	D	H	V	V	S	G	N	L	V	T	P
SS	C	C	E	E	C	C	C	C	C	C	E	E	E	E	E	E	C	C	C	C
AA	L	P	V	I	L	G	H	E	A	A	G	I	V	E	S	V	G	E	G	V
SS	C	C	C	C	C	C	C	E	E	E	E	C	C	C	C	C	C	C	C	C
AA	T	T	V	K	P	G	D	K	V	I	P	L	F	T	P	Q	C	G	K	C
SS	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
AA	R	V	C	K	N	P	E	S	N	Y	C	L	K	N	D	L	G	N	P	R
SS	C	C	C	C	C	C	C	C	C	C	C	C	C	C	E	E	E	E	E	E
AA	G	T	L	Q	D	G	T	R	R	F	T	C	R	G	K	P	I	H	H	F
SS	E	C	C	C	C	C	C	C	C	E	E	E	E	C	C	E	E	E	E	C
AA	L	G	T	S	T	F	S	Q	Y	T	V	V	D	E	N	A	V	A	K	I
SS	C	C	C	C	C	C	C	E	E	E	E	E	C	C	C	C	C	C	C	C
AA	D	A	A	S	P	L	E	K	V	C	L	I	G	C	G	F	S	T	G	Y
SS	C	C	E	E	E	C	C	C	C	C	C	C	C	E	E	E	E	E	C	C
AA	G	S	A	V	N	V	A	K	V	T	P	G	S	T	C	A	V	F	G	L
SS	C	H	H	H	H	H	H	H	H	H	H	H	H	C	C	C	C	E	E	E
AA	G	G	V	G	L	S	A	V	M	G	C	K	A	A	G	A	A	R	I	I
SS	E	E	C	C	C	C	C	H	H	H	H	H	H	H	C	C	C	C	C	C
AA	A	V	D	I	N	K	D	K	F	A	K	A	K	E	L	G	A	T	E	C
SS	E	E	E	C	C	C	C	C	C	C	C	C	E	E	E	E	E	E	C	C
AA	L	S	I	N	P	M	L	L	L	T	G	R	T	W	K	G	A	V	Y	G
SS	C	C	C	C	C	C	C	H	H	H	H	H	H	H	H	H	C	C	C	C
AA	G	F	K	S	K	E	G	I	P	K	L	V	A	D	F	M	A	K	K	F
SS	C	C	C	C	C	C	C	C	C	C	C	C	C	H	H	H	H	H	H	H
AA	S	L	D	A	L	I	T	H	V	L	P	F	E	K	I	N	E	G	F	D
SS	H	H	H	C	C	C	C	E	E	E	E	E	E	C						
AA	L	L	H	S	G	K	S	I	C	T	V	L	T	F						

In Table 2-4, where the row starts with an **SS** *(secondary structure)* each cell in that row contains the classified secondary structure element for the amino acid residue presented in the cell below, the cell value either contains (h) helix, (e) strand or (c) coil, these are the classified secondary elements of the protein. Where the row starts with **AA** *(amino acid)*, each cell in that row contains an amino acid residue for the protein sequence 1THB as shown in Figure 2-4.

2.2.3 Tertiary structure

The three-dimensional (3D) structure of a protein is a key determinant of its biological function. A proteins tertiary structure also known as 3D structure is the next step once secondary structure elements are identified (Kurgan and Homaeian 2006). A protein folds its linear primary sequence of amino acids into its native state by using minimum energy (Eidhammer I 2005); the folding process packs the secondary structure elements into tightly packed pre-arranged structure (Cohen and Kuntz 1987) - represented in Figure 2-7.

The native state of a protein is its properly folded form, which gives it function (Cai, Liu et al. 2001; Eidhammer I 2005). Finding the native state using classification methods is one of the biggest challenges in protein structure prediction as a whole because the computational search space to find proteins native state can run into the millions of possible conformations (Chou and Zhang 1995; Chou 2005). An example of large search spaces, a peptide sequence of just 10 amino acid residues has 3628800 possibilities (10!) to search through to find its native state. Proteins with larger number of residues will have hundreds of millions or billions conformational possibilities, which is impossible to search solely through without some mechanism of filtering a search space by inputting as much as known information of the protein as possible.

The importance of knowing the structure of a protein, is so attempts can be made to design drug molecules to bind to it and block or express more of its activity function which could ultimately help battle against certain diseases (Goodsell 2010).

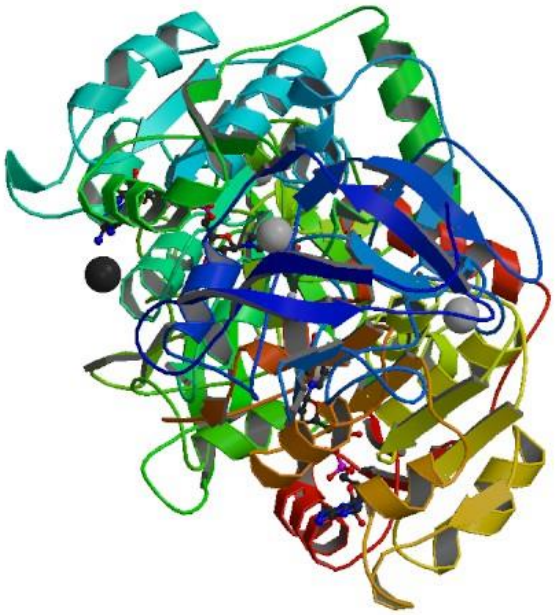


Figure 2-7 3D Crystallization of Human Beta3 Alcohol Dehydrogenase PDB ID 1HTB (Berman 2007)

2.2.4 Current experimental procedures to determine protein structures

Nuclear magnetic resonance (NMR) and x-ray crystallography are the two main methods that are currently applied to the study of three-dimensional structures of proteins, which is at atomic resolution (Sneath 1966; Mielke and Krishnan 2003; Chou 2005). NMR is a method that allows the determination of three-dimensional structures of proteins in the solution phase of experimental biology (Mielke and Krishnan 2003). X-ray crystallography is a method being used to determine the arrangement of atoms within a crystal, in which beam of x-rays strikes a crystal and causes the beam of light to spread into many specific directions. Proteins are crystallised to provide a diffraction pattern when exposed to X-rays beams and allows the production of multiple copies of the same protein (Hobohm and Sander 1994). The experimental primary, secondary and tertiary structures information can be obtained from main data resources such as the PDB, UniProt and SwissProt. The PDB data resource contains 87279 experimentally verified protein structures each having a primary, secondary and tertiary structures information. This figure can be further broken down by the methods of different experimental procedures, x-ray 75506 and NMR 9608 (Berman 2007) as of October 2012.

2.2.5 Transition from secondary structures to structural classes

The meaning of secondary structure classification is to classify the local secondary structures of proteins based only on knowledge of their primary amino acid sequence. The local secondary structures are α-helices and β-sheets, Figure 2-5 and Figure 2-6, respectively. A classification consists of assigning regions of the amino acid sequence space as probable alpha helices, beta strands or coils (Yang, Peng et al. 2010) as mentioned in section 2.2.2. Protein structural classification is based on labelling what the protein sequence main secondary structure element is i.e. classifying the majority secondary structural element found in a protein. For example, protein 1HTB structural class is all-α. An all-α protein is a class of structural domains in which the majority secondary structure content is composed of α-helices, possibly with few isolated β-sheets. Reasons for classifying structural classes are given in section 2.2.6. The ever-increasing gap between the output of protein sequences and structural verification of them creates difficulty closing the gap. Thus, research has turned to obtaining additional knowledge about protein from the structural space (Ahmadi Adl, Nowzari-Dalini et al. 2012; Xia, Ge et al. 2012). Fully annotated proteins are limited in terms of number of protein sequences in the NCBI RefSeq database (Pruitt, Tatusova et al. 2009; Pruitt, Tatusova et al. 2012). As of February 2013, there are 88,170 fully annotated proteins and 27,834,581 non-annotated protein sequences. This large gap places more importance on the needs for automated and precise sequence-based prediction methods. Precise prediction of structural classes of proteins has been proven valuable for other proteomic studies such as the prediction of protein secondary and tertiary structures (Deleage and Roux 1987) and (Costantini, Colonna et al. 2007)

2.2.6 Structural classes

A prior knowledge of structural classes of proteins has become quite useful from both an experimental and theoretical point of view (Levitt and Chothia 1976; Luo, Feng et al. 2002; Gromiha and Selvaraj 2004; Chou 2005; Yu, Sun et al. 2007). Knowledge of structural classes of proteins is important in many respects:-

- The knowledge of the structural class of a protein reduces the conformational search space during the search of the tertiary structure (Cohen and Kuntz 1987; Bahar, Atilgan et al. 1997) as it presents a description of the proteins overall folding process (Wei-Shu Bu 1999).

- Classification of structural classes of proteins enables the identification of common structural patterns of proteins and it shows that the arrangement of secondary structure elements along the sequence relates to three-dimensional properties of the protein (Levitt and Chothia 1976).
- Secondary structure determination from the primary sequence is improved by incorporating knowledge of structural class (Cohen and Kuntz 1987; Cohen and Kuntz 1987; Deleage and Roux 1987; Chou 1989; Deleage and Dixon 1989; Kneller, Cohen et al. 1990; Muggleton, King et al. 1992) and vice versa, in some studies the predicted secondary structure information has helped with the classification of structural classes of proteins (Yang, Peng et al. 2010)
- Reduce the gap between known structural class domains and the unavailability of experimental protein structure information, which is used to assign the structural class for the majority of known protein sequences (Ke Chen 2008).

The knowledge of the structural classes of proteins is also a useful property applicable to the wider area of proteomics. Such as protein localisation (where it resides within an organism cells) and what type the protein is i.e. enzyme or non-enzyme (Nishikawa and Ooi 1982; Kidera, Konishi et al. 1985; Chou 1995; Chou and Zhang 1995; Eisenhaber, Frömmel et al. 1996; Chou 2005; Kurgan and Homaeian 2006; Kawashima, Pokarowski et al. 2008). Levit and Chothia developed the original concept of protein structural classes (Levitt and Chothia 1976). From their work on globular proteins, they saw that protein structures naturally grouped into four main structural classes based on the gross amount of secondary structure elements found in tertiary structures (Levitt and Chothia 1976). They devised a system that categorised proteins into one of the following four classes.

a) All-α: proteins with only small amount of beta-strands Figure 2-8

b) All-β: proteins with only small amount of alpha-helices Figure 2-9

c) α/β: proteins that include alpha-helices and beta-strands, where beta-strands are mostly parallel Figure 2-10

d) α+β: proteins with both alpha-helices and beta-strands, where beta-strands are mostly anti-parallel Figure 2-11

This classification is based upon majority secondary structure content (Zhang and Chou 1992) present in a protein.

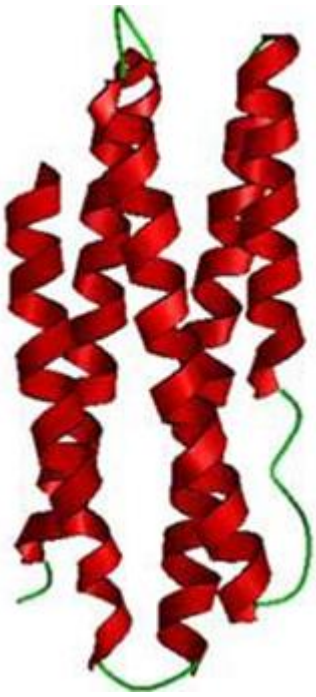


Figure 2-8 Ribbon representation of all-α structural class (Chou 2005)

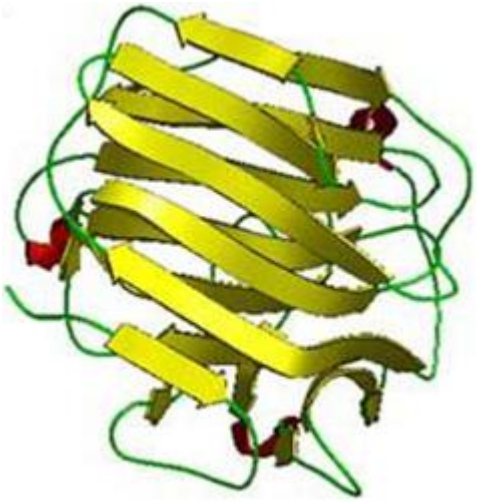


Figure 2-9 Ribbon representation of all-β structural class (Chou 2005)

Figure 2-10 Ribbon representation of α/β structural class (Chou 2005)

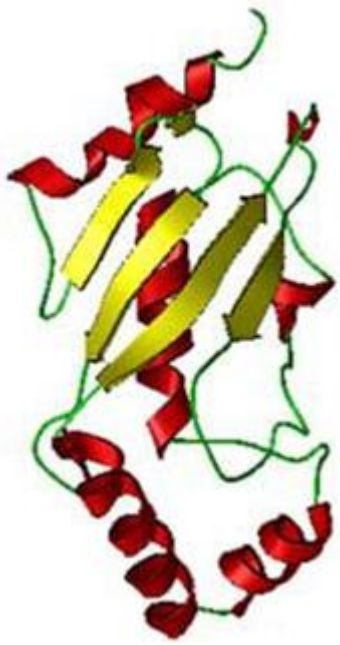


Figure 2-11 Ribbon representation of α+β structural class (Chou 2005)

α-helices and β-strands are the structural elements found in secondary structure composition. The difference between α/β and α+β structural classes of proteins are how the α-helices and β-strands are arranged in the protein. In the α/β structural class, the α-helices commonly separated from the β-strands and alternate more frequently than α+β structural class. Compared to the α+β structural class the β-strands are usually interspersed (Chou and Zhang

1995). Majority of proteins are categorised into these main four classes, however, in addition to the main four groups there are several other structural classes in which very small numbers of proteins exists (Murzin, Brenner et al. 1995):-

1. Multi-domain proteins (alpha and beta) - folds consisting of two or more domains which belong to different structural classes
2. Membrane and cell surface proteins
3. Small proteins
4. Coiled coil proteins
5. Low resolution protein structures
6. Peptides
7. Designed proteins

How does protein structure determine function? The tertiary structure of a protein infers the biological function of the protein (Yang, Peng et al. 2010). An analogy to protein structure linked to protein function is the key and lock, the correct key unlocks the correct door; the correct shape (structure) allows the inferring of function (Eidhammer I 2005). Knowing the structural classes of proteins is an important aspect in tertiary structure classification as the structural class of a protein presents an intuitive description of its overall folding process. With many millions of conformations a protein can fold into the restriction the structural class imposes is a reduction in the search space to find the single conformation state the protein folds into and has high impact on its tertiary structure classification (Chou and Zhang 1995)

Several studies have proposed definitions of structural class thresholds; these have been revised a few times since the conception of structural classes of protein back in the 1970's (Chothia 1976; Levitt and Chothia 1976). Table 2-5 contains thresholds of α and β structural elements, which were used to determine which structural class a proteins is categorised into. The difference between each system is the amount of helices found in all-β proteins and the amount of strands found in all-α protein (Kurgan and Homaeian 2006).

Table 2-5 Structural class thresholds (Kurgan and Homaeian 2006)

Structural class	Helix (α) threshold	Strand (β) threshold	Reference
all-α	> 15%	< 10%	(Nakashima, Nishikawa et al. 1986)
all-β	< 15%	< 10%	
α+β	> 15%	> 10%	
α/β	> 15%	> 10%	
all-α	≥ 40%	≤ 5%	(Chou 1995)
all-β	≤ 5%	≥ 40%	
α+β	≥ 15%	≥ 15%	
α/β	≥ 15%	≥ 15%	
all-α	> 15%	< 10%	(Eisenhaber, Frömmel et al. 1996; Eisenhaber, Imperiale et al. 1996)
all-β	< 15%	> 10%	
α+β	> 15%	> 10%	
all-α	Manually classified (SCOP database)		(Murzin, Brenner et al. 1995)
all-β			
α+β			
α/β			

However, the thresholds system has now been replaced by a manual classification method named the Structural Classification of Proteins (SCOP), which is a manual method replacing the various thresholds as it became too rigid and other times too relaxed to categorise proteins into appropriate structural classes. The assignment of the structural class of proteins is now performed manually. The SCOP databases stores the structural classes of proteins that have been manually verified (Murzin, Brenner et al. 1995). SCOP has become one of the main data resources used in protein structural class research. The SCOP database has manually determined the structural class of many of the proteins from the PDB that have been experimentally verified (Murzin, Brenner et al. 1995). The PDB uses the assignment of structural classes of proteins done by SCOP into its own databank, not all proteins from the PDB are assigned a structural class yet, as SCOP is still going through PDB data and assigning each experimentally verified protein a structural class. Table 2-6 contains the number of PDB and SCOP entries as of February 2013; the column named "No. of proteins…" contains the number proteins that have been experimentally classified by PDB and that have had a structural class assigned by SCOP. The protein datasets that will be used for the thesis analyses will come from the PDB and SCOP databases as they have experimentally determined protein structures where its structural classes are known.

Table 2-6 Classification of proteins into structural classes

Structural class	No. of proteins where class information is known	Reference
All alpha proteins	7627	(Goodsell 2010)
All beta proteins	10672	
Alpha and beta proteins (α/β)	11965	
Alpha and beta proteins (α+β)	11053	
Small proteins	2282	
Multi-domain proteins (alpha and beta)	1199	
Peptides	773	
Other	1592	
Total	47163	

The ASTRAL compendium for sequence and structure analysis provides databases and tools used in the research and analysis of protein structures and their sequences, the sequences partly derived from SCOP database. The other part of the data is derived from the coordinate files maintained and distributed by the PDB (Chandonia, Hon et al. 2004). The ASTRAL compendium is useful to extract large number of protein's amino acid sequences whose structural classes have been determined experimentally already. These sequences can form datasets to be used in bioinformatics studies.

2.3 Bioinformatics for prediction of structural classes of proteins

Developments in molecular and structural biology during the last three decades, along with the development of large-scale genome technologies and the need to study complex biological systems, have led to the exponential growth and development of biological data produced. Bioinformatics is the application of computing to the organisation and analysis of biological data. The consequence is that computers are being used to collect, store and analyse biological data. Bioinformatics is a multidisciplinary research area that is the interface between the biological and computational sciences with applied mathematics and statistics. The goal of bioinformatics is to uncover the hidden biological information from the large set of data and obtain a richer insight into the essential biology of organisms. There is a huge need for computationally developed tools and methodologies aiming to manage, control and analyse the large amount of biological information available, in order to derive meaningful information hidden in the data (Popov, Nenov et al. 2009).

2.3.1 Real world bioinformatics applications

Since the final version of the human genome project completed in April 2003 it has had a big impact on current biomedical research and clinical medicine research. It has allowed scientists unprecedented access to biological data to search for genes directly related to different disease, to try to understand the molecular fabric of the disease. Finding new knowledge within the data brings with it better understanding, which can translate into better treatments and cures. The finding of new knowledge is the key to finding new drug targets, as currently there are around 500 proteins as drugs target for many different illnesses and diseases. Drugs that are highly specific to a target protein have fewer side effects (Overington, Al-Lazikani et al. 2006; Karadaghi 2012; Li, Huang et al. 2012; Li, Hu et al. 2012).

2.3.2 Bioinformatics and Proteomics

Proteome is the collection of all the proteins in any given organism and proteomics is the study of the protein structures and functions found in the proteome. Proteomics methods are essential for studying protein activity, protein expression, protein regulation and protein modifications. The most important study in proteomics is the determination of the protein's tertiary structures as knowing the tertiary structure helps deduce the function of a protein. Bioinformatics is an integral part of proteomics research, more so, the informatics in proteomics is progressively increasing because of the arrival of high-throughput computational methods relying on very powerful data analysis, which needs powerful computing and reliable storage systems (Popov, Nenov et al. 2009).

2.3.3 Bioinformatics for the prediction for structural classes of proteins by using sequence information

One of the most abundant protein information available is its amino acid sequence; and these have been put into many datasets that have been constructed for the representations of structural classes of proteins (Kurgan and Homaeian 2006). Using these datasets, sequence driven features can be extracted that turn a varying length of protein amino acid sequence into a fixed length vector. Sequence driven features are widely used because many different protein information can be extracted from a protein sequence. Using sequence driven features, a predictive model can be developed by using classification tools. To assess the predictive model the model is evaluated using test procedures where the accuracy of predictive model is obtained. There are many different variations to each of these areas, which affect the final overall predictive accuracy of the selected dataset. All these different

areas put together is called a bioinformatics approach for prediction of structural class of proteins. It has been shown that the dataset size, class definitions and test procedures are important factors to consider for obtaining reliable predictive accuracy (Eisenhaber, Frömmel et al. 1996; Eisenhaber, Imperiale et al. 1996).

2.3.4 Data resources

Data resources protein data bank (PDB) and structural classification of proteins (SCOP) are the main resources where protein amino acid sequences are obtained and constructed. These datasets contain the collection of protein amino acid sequences used to train and test predictive models. The datasets used or developed must be representative of the prediction of protein structural classes, taking into consideration of the homology level and sample sizes of each structural class as this impacts on predictive accuracy (Kurgan and Homaeian 2006). There have been many datasets developed for the prediction of protein structural classes, which are mainly derived from SCOP database; these are listed in section 2.3.5.

The PDB database contains information about experimentally determined structures of proteins; it is the major resource for fully annotated proteins including structural class information, which is derived from the Structural Classification of Proteins (SCOP) database – more information about SCOP is given in the next section. Popular datasets constructed using PDB are listed in Table 2-7.

SCOP is a manually annotated database and has been regarded as the most accurate classification of structural classes of proteins (Chandonia, Hon et al. 2004). The SCOP database contains proteins that are manually curated, annotated and classified into the structural classes providing information about folds evolutionary relationships, which can then be used in numerous protein structure related studies (Kurgan and Homaeian 2006). The current version of the SCOP database, v. 1.75, includes eleven structural classes, with the four major classes (all-α, all-β, α/β and α+β) covering approximately 90% of the entries in PDB (Murzin, Brenner et al. 1995; Xia, Ge et al. 2012).

2.3.5 Datasets constructed using PDB and SCOP

These are some of the datasets constructed using SCOP database and used in the development of prediction models for structural classes of proteins, to name some are :-

- 1189 (Wang and Yuan 2000), the 1189 dataset was processed and filtered using the SCOP (version, 1.67) and PDB release as of February 2005.
- 25PDB (Hobohm and Sander 1994), 15 times smaller than the PDB database and includes only high quality non-homologous proteins as of February 2005, 2340 sequences and domains were extracted.
- 204 (Chou 1999) 04 non-homologous proteins extracted from the PDB.
- 359 homology unknown (Chou and Maggiora 1998), dataset 359 is relatively small; it was the most extensively used in the past studies. It includes 359 highly homologous domains and sequences.

Table 2-7 Datasets constructed using PDB and SCOP

Dataset Name	Structural Class				References
	All-α	All-β	α/β	α+β	
1189	223	294	334	241	(Wang and Yuan 2000)
25PDB	443	443	346	441	(Hobohm and Sander 1994)
204	52	61	45	46	(Chou 1999)
359	82	85	99	93	(Chou and Maggiora 1998)

The ASTRAL database (Chandonia, Hon et al. 2004) is the collection of protein sequences from the SCOP database. The ASTRAL database is a very important data resource as it contains the largest set of sequences that have already been manually assigned its structural classes. The ASTRAL database contains a feature that allows the construction of protein datasets representing structural classes at user defined sequence homology levels.

2.3.5.1 Class Architecture Topology and Homologous (CATH)

The Class Architecture Topology and Homologous (CATH) database is a semi-automated protein structural class database. Protein domains from single and multi-domain protein structures of the PDB are classified at four levels (Orengo, Michie et al. 1997), as:-

(1) Class – protein structures are classified according to their secondary structure composition (mostly alpha, mostly beta, mixed alpha/beta or few secondary structures).

(2) Architecture - protein structures are classified according to their overall shape as determined by the orientations of the secondary structures in 3D space but ignore the connectivity between them.

(3) Topology (fold family) - protein structures are grouped into fold groups at this level depending on both the overall shape and connectivity of the secondary structures.

(4) Homologous superfamily - this level groups together protein domains which are thought to share a common ancestor and can therefore be described as homologous

The difference between CATH and SCOP databases, CATH does not differentiate between α/β and α+ β structural class, these two structural classes are treated as one mixed αβ structural class. The CATH methods provide a much quicker classification of a proteins structural class, but it lacks where the SCOP manual methods are more refined and has bigger database covering more structural classes in (Csaba, Birzele et al. 2009).

2.3.6 Sequence driven features for protein representation

Proteins amino acid sequence come in different lengths, from just a few amino acid residues to sequence containing thousands of amino acid residues. The reason why protein amino acid sequences come in different lengths is that every protein come in a wide variety of physical shapes, sizes and functions and has a unique sequence of amino acid residues (Lamond 2002). Many thousands of proteins have been experimentally verified each with its own particular amino acid sequence. To be used effectively in computational methods, protein sequences need to be transformed from a string amino acid residues into a fixed length vector (Karchin, Karplus et al. 2002) that can be inputted into classification and clustering classifiers. These fixed length vectors are also known as sequence driven features. Sequence driven features are widely used in proteomic studies because many different protein information can be derived from the protein amino acid sequence, which is an important part of protein prediction modelling (Chou 2005; Marsolo and Parthasarathy 2006). All proteins in the PDB have an amino acid sequence which makes the amino acid sequence one of the most widely available protein property (Ahmadi Adl, Nowzari-Dalini et al. 2012). Although there are various types of sequence, driven features presented in the literature they can be mainly categorised into 9 different types (1) amino acid composition, (2) dipeptide composition, (3) autocorrelation, (5) composition, (6) transition, (7) distribution, (8) sequence-order and (9) pseudo amino acid composition , the details of each feature groups are given in chapter 4.

Sequence driven features have been used in many prediction of protein structural classes studies (Chou 1989; Cid, Bunster et al. 1992; Eisenhaber, Frömmel et al. 1996; Bahar, Atilgan et al. 1997; Dubchak, Muchnik et al. 1999; Luo, Feng et al. 2002; Xiao, Shao et al. 2006; Ding, Zhang et al. 2007; Zhang, Ding et al. 2008) and other proteomic properties such as protein-protein interactions (Bock and Gough 2001; Lo, Cai et al. 2005), subcellular locations (Cai, Liu

et al. 2001; Chou and Cai 2004), peptides containing specific properties (Cui, Han et al. 2007) and functional classes (Cai, Han et al. 2003).

2.3.7 Amino acid indices

Amino acid indices have been investigated through many experimental and theoretical studies since the early sixties (Kawashima, Pokarowski et al. 2008). Each property can be represented by fixed set of twenty descriptor values, which are known as an amino acid index representing a certain physiochemical or biochemical property of a protein, such as alpha and turn propensities, beta propensity, composition, hydrophobicity and physicochemical properties (Tomii and Kanehisa 1996).

Such amino acid indices are housed in the Amino Acid Index database (AAIndex1) (Kawashima, Pokarowski et al. 2008). Amino acid indices are used in current sequence-driven-features such as autocorrelation and pseudo amino acid composition. However, the utilisations of these amino acid indices have been limited to a small subset of the available numbers of indices. Amino acid indices of different physicochemical and biochemical properties have been extensively used in various bioinformatics studies, such as predicting protein secondary structures (Kazemian, Moshiri et al. 2007), trans membrane sequences (Zhao and London 2006) and surface (Nishikawa and Ooi 1980; Nishikawa and Ooi 1986). Amino acid indices have also been used to derive new indices (Huang, Kawashima et al. 2007) and numerical representations of amino acid residues for establishing the structure of proteins (Georgiev 2009). In bioinformatics analysis, the correct selection of amino acid indices representing their biological significance is essential for efficient representation of the feature against the problem (Saha, Maulik et al. 2011).

2.3.8 Predictive models

Predictive models are the algorithms used for classification of proteins. A classification involves separating protein structural class datasets into training and testing sets. Each protein sample in the training set contains one target value (i.e. class labels) and several attributes (i.e. the features or observed variables of the sample). The classification algorithm builds a predictive model based on the training set of data samples, which is used to best classify the testing set of protein samples. Algorithms that have been used in the field are-

- k-nearest Neighbour (Cover and Hart 1967)

- Vector decomposition (Eisenhaber, Frömmel et al. 1996; Eisenhaber, Imperiale et al. 1996)
- Geometric classification (Chou and Maggiora 1998)
- Component coupled geometric classification (Wei-Shu Bu 1999)
- Bayesian classification (Wang and Yuan 2000)
- Discriminant analysis (Luo, Feng et al. 2002)
- Information discrepancy based classification (Jin, Fang et al. 2003)
- Intimate sorting classification (Chou and Cai 2004)
- Support vector machine (Cortes and Vapnik 1995)

The most widely used classification methods used in literature are support vector machine (Cai, Liu et al. 2001; Ding and Dubchak 2001; Cai, Liu et al. 2002; Cai, Liu et al. 2003; Markowetz, Edler et al. 2003; Isik, Yanikoglu et al. 2004; Chen, Zhou et al. 2006; Zhang and Ding 2007; Anand, Pugalenthi et al. 2008; Chen, Chen et al. 2008; Melvin, Weston et al. 2008; Jian-Ding, San-Hua et al. 2009; Wang, Wang et al. 2011) and k nearest neighbour (KNN) classifier (Grassmann, Reczko et al. 1999; Zhang, Wang et al. 2005; Chen, Chen et al. 2008; Li, Lin et al. 2008; Melvin, Weston et al. 2008; Zhang, Ding et al. 2008; Liu, Zheng et al. 2009; Hernández-Rodríguez, Martínez-Trinidad et al. 2010). SVM and KNN are popular algorithms when there are little or no prior knowledge about the structure or distribution of the datasets. K-nearest neighbour algorithm (KNN) is categorised as a supervised and nonparametric classification algorithm for which classifying testing data point based on the nearest training data points in the feature space. The k-nearest neighbour algorithm is one of the simplest machine-learning algorithms as testing data point is classified by a majority vote of its neighbour's; with the testing data point being assigned to the class most common amongst its k-nearest-neighbours (k is a positive number). If k=1, then the testing data point is assigned to the class assigned to its first nearest neighbour, similarly if k=3 the testing data point is assigned to its third nearest neighbour. Detailed information regarding SVM and KNN are given in chapter 3 (Cover and Hart 1967; Chen, Chen et al. 2008; Hmeidi, Hawashin et al. 2008). SVM is a supervised and parametric machine-learning algorithm for data classification. The principle of SVM is to seek an optimum hyperplane for data classification. SVM uses different kernel functions to map the input data to a higher dimensional space where it seeks a hyperplane to separate the training protein samples by their classes, the most common ones are linear, polynomial, radial basis function and sigmoid. The input instances which are close

to the hyperplane are called support vectors and are crucial for training (Cortes and Vapnik 1995; Cai, Liu et al. 2001; Chang and Lin 2001; Ding and Dubchak 2001). SVM is very sensitive to the parameter selection and requirecareful selection, as the incorrect values will result in unreliable results. These parameters are *C* - the penalty factor and gamma parameter for the kernel which determines the shape of which data points are projected into a higher dimension (Chang and Lin 2001) and there is a process called model selection where the two main parameters are selected based on a grid search of optimal values based on the dataset.

2.3.9 Assessment of the predictive models (test procedures)

Classification methods have deal with assigning predicted labels to test samples. To evaluate the prediction accuracy of predicted test samples, test procedures should be used to assess the generalisation ability of the method. In order to evaluate predictive accuracy of the classification method, the test procedure partitions the dataset into training and testing subsets. Different test procedure partitions the dataset into training and testing subs sets differently. Whichever test procedure is used to partition the dataset, the classification algorithm will use the training subset to build the predictive model and use the testing subsets to test the predictive model on and then they derive the accuracy based on correctly predicted samples (Hotta, Kiyasu et al. 2004; Nigsch, Bender et al. 2006; Shen and Chou 2009).

There are three commonly used test procedures used in protein structural class predictions, they are, (1) resubstitution (2) n-fold and (3) leave-one-out (LOO) (Gu and Chen 2009; Liu, Zheng et al. 2009; Sahu, Panda et al. 2009; Shen and Chou 2009; Yang, Peng et al. 2009; Sakar, Kursun et al. 2010; Yang, Peng et al. 2010; Wang, Wang et al. 2011; Ahmadi Adl, Nowzari-Dalini et al. 2012; Ding, Zhang et al. 2012). LOO is also known as jackknife in the wider field but it is used interchangeably in the literature, within this thesis it is referred to as leave-one-out. Each of the test procedures partitions the datasets differently (sub sampling) which has been shown to yield different predictive accuracy rates (Gu and Chen 2009; Liu, Zheng et al. 2009; Sahu, Panda et al. 2009; Shen and Chou 2009; Yang, Peng et al. 2009; Sakar, Kursun et al. 2010; Yang, Peng et al. 2010; Wang, Wang et al. 2011; Ahmadi Adl, Nowzari-Dalini et al. 2012; Ding, Zhang et al. 2012).

Test procedure tests the predictive model on the testing data test. Resubstitution partitions the dataset where a proportion of the data samples appear in both training and testing sub datasets. In effect, testing sample is predicted using its own information, which leads to high

accuracies (Chou and Maggiora 1998; Zhang, Wang et al. 2005). Although it is commonly recognised that resubstitution test procedure leads to high accuracies it is not a reliable way of assessment (Kurgan and Homaeian 2006). However, it is still commonly used in studies as shown in Table 2-8 but results cannot be reliable or generalised.

In LOO test procedure, one sample is taken out of dataset for testing and trained upon using remaining training samples. The *N* samples-1 where N = total number of sample minus one (i.e. leave one out) is put into the training subset and the remaining sample is put into the test subset. Leave-one-out is perceived to be more rigorous and reliable test procedure but computationally demanding one as it evaluates each sample against the *N* samples-1 training subset (Kurgan and Homaeian 2006). Leave-one-out is suitable for small datasets around (i.e. 2000 samples) but for larger datasets (i.e. over 5000 samples), it becomes too computationally resourceful.

N-fold cross validation partitions the dataset into n-folds, the number folds present in the training subset is *n*-1 fold and remaining fold is used in the testing subset. Common *n*-fold values are 5 and 10, meaning that the dataset is divided into 5 or 10 subsets, each of which contains more or less the same number of samples from each class (Kurgan and Homaeian 2006). N-fold is considered as a very fast and reliable test procedure as it evaluates a larger test sub-sets on a smaller training subset.

Independent-sets test procedure is independent from the training set validation of the predictive model, i.e. it consists of a separate dataset. Independent-sets test procedure is not a widely used evaluation method in the prediction of protein structural classes, but it is gaining popularity (Mizianty and Kurgan 2009; Ding, Zhang et al. 2012; Karadaghi 2012). Training and testing datasets are made up of two distinct datasets with no sub sampling involved (Mizianty and Kurgan 2009).

The importance of test procedures is to evaluate classification accuracy and generalization abilities of the classification algorithms. Further details of leave-one-out, n-fold and independent-sets are presented in chapter 3.

2.3.10 Sequence homology

One of the factors that play an important role in the prediction of structural classes as well as functional properties of the proteins is sequence homology. Sequence homology refers to the

situation where protein sequences are similar because they have a common evolutionary origin. Protein datasets come in of varying homology levels between 20% to over 90% (Whitfield, Pruess et al. 2006). Protein datasets characterised by homology levels between 20–30% sequences similarity is called twilight zone proteins (Rost 1999). Between 30-40% is the normal accepted homology rate within datasets. Anything higher than 50% is considered highly homologous dataset whereby all sequences on average is 50% similar with each other.

Sequence homology affects prediction accuracy where the computational method relies on the same protein dataset for training and testing the predictive classifier. Highly homologous protein datasets tend to result in higher prediction accuracies because the sequence similarities are high the test procedure relies on the same high homology dataset for testing and training the predictive model. Example, highly homologous dataset at 60%, the predictive classifier will train and test on a dataset that is 60% similar which will result in a higher predictive accuracy because prior knowledge is known of the test set of samples from the training set of samples (Nakashima, Nishikawa et al. 1986) (Wang and Yuan 2000) as shown in Table 2-13. Many published studies analyse using homologous datasets that produces high predictive accuracies and should be considered with caution, as results may not be reliable at such levels. Current prediction accuracies reported in various studies are listed in Table 2-13.

An example of highly homologous dataset is denoted as 359 (Chou and Maggiora 1998) and it was extensively used in the past studies. The results obtained using this dataset range from 84.1% to 97% using leave-one-out test procedure and 94.3% to 100% by using resubstitution test procedure (Chou and Maggiora 1998; Wei-Shu Bu 1999; Cai, Liu et al. 2003; Jin, Fang et al. 2003; Kurgan and Homaeian 2006). Jaroszewski et al showed that among the 359 sequences only, over 100 sequences out of 359 are identical (Li, Jaroszewski et al. 2001; Li, Jaroszewski et al. 2002). These results reveal that high homology can affect result and overestimate prediction accuracy.

2.3.11 Feature selection

Feature selection (FS) is the process of finding a smaller subset of features that is representative of all the features from a dataset. One of the issues in proteomic studies is that usually the numbers of features are significantly larger than the number of protein samples, which can include a high amount of noisy data. Most often, datasets have many thousands of features that represent its samples, not all the features are relevant to the sample or the

whole dataset, this is where FS is becoming an important process because with FS application it is able to remove irrelevant features (noisy ones) from a dataset (Peng, Long et al. 2005; Liang and Zhao 2006; Lin, Chun-Yuan et al. 2007).

The motivation for feature selection is to find the most important features from the dataset without over fitting to improve predicative accuracy using classification methods. This will enable a more efficient predictive model by removing irrelevant features (Saeys, Inza et al. 2007). This also helps speed up the computational analysis time as the predictive classifier deals with a smaller subset of the original feature set after feature selection. Feature selection tools have been used in many bioinformatics studies ranging from gene selection/expression data (Ding, Peng et al. 2003; Li, Huang et al. 2012), protein-protein interaction network (Li, Huang et al. 2012), protein sub-nuclear location (Sakar, Kursun et al. 2010) and including protein structural classes (Chen, Lu et al. 2009; Jahandideh, Hoseini et al. 2009).

Many feature selection methods have been presented in the literature. These methods can however be described in three main groups; (1) filter, (2) wrapper and (2) embedded (Saeys, Inza et al. 2007). The filter technique looks at the basic properties of each feature and scores each one independently or as mutual set of feature as illustrated in Figure 2-13. The lower the feature score the less important the feature is. The wrapper technique combines the feature selection method with a classifier; it selects the optimal features, which is dependent on the selection of classifier, illustrated in Figure 2-14. The embedded techniques continually searches for an optimal set of features and analyses them for the best accuracy, this search is built into the specific classifier and will make hypotheses as it learns which features a more or less representative, this is illustrated in Figure 2-15.

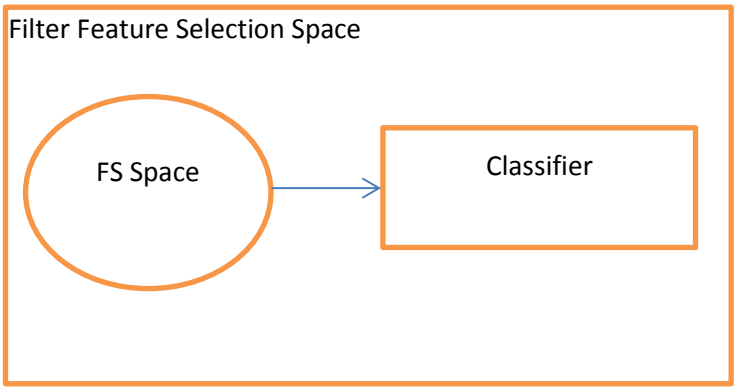


Figure 2-12 Filter Feature Selection Space (Saeys, Inza et al. 2007)

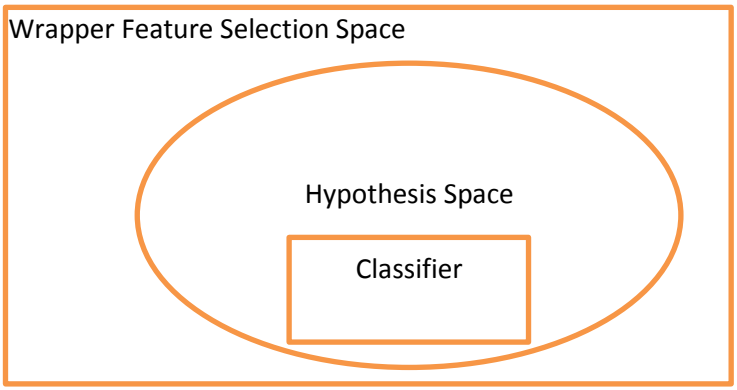


Figure 2-13 Wrapper Feature Selection Space (Saeys, Inza et al. 2007)

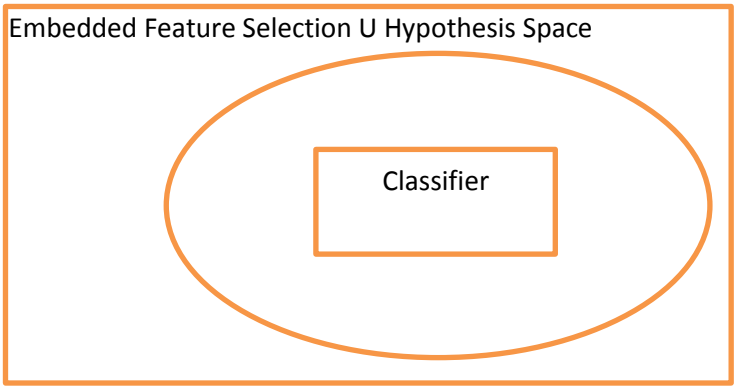


Figure 2-14 Embedded Feature Selection (Saeys, Inza et al. 2007)

The two most popular feature selection methods are F-score and minimal-redundancy-maximal-relevance (mRMR).

F-score

F-score is a simple filter technique independent of any classification methods, which measures the discrimination of two sets of features. A disadvantage of F-score it does not reveal mutual information among features, as each feature (regardless if they are logically grouped) is independently scored (Xu, Liu et al. 2008). Although F-score is independent of any classification algorithm it is widely used in combination with SVM and Random Forest (RF) by selecting features with high F-scores and then applying either SVM or Random Forest classification (Chang and Lin 2001).

Minimal-redundancy-maximal-relevance (mRMR)

The mRMR feature selection method is a multivariate filter method which takes into consideration the mutual information, correlation, distance/similarity values of the whole feature space to select relevant features (Peng, Long et al. 2005). mRMR is independent of any prediction model (classification method). The maximum relevance looks for features that add the most significant value to the target classification and the minimum redundancy section works to eliminate the features whose prediction capability has already been included by the selected features i.e. it has removed the noisy features (Peng, Long et al. 2005; Li, Lin et al. 2008; Zhang, Ding et al. 2008).

2.3.12 Current prediction accuracies

Table 2-8 contains current prediction accuracies for the benched mark datasets 25PDB and 1189. This table highlights the studies that have used sequence driven features of many different types with varying feature sizes and classification algorithm. The highest accuracy obtained for the 25PDB dataset is 62.7% (Kurgan and Chen 2007) and for the 1189 dataset is 67.6% (Ke Chen 2008) using leave-one-out test procedure.

Table 2-8 Current prediction accuracies and sequence representation for four class protein structural class predictions (Kurgan and Homaeian 2006; Yang, Peng et al. 2010)

Classification Algorithm	Sequence Representation		Classes	Dataset			Classification Accuracy		
	Feature Name	Feature Size		Size	Homology	Domains	Resubstitution	Leave-one-out	Reference
Vector decomposition	Amino Acid Composition	20	3	260	Unknown	No	60.8%	57.7%	(Eisenhaber, Frömmel et al. 1996; Eisenhaber, Imperiale et al. 1996)
				471	Unknown	No	58.2%	57.3%	
Geometric classification	Amino Acid Composition	20	4 classes (SCOP)	359	Unknown, but homologous	Yes	94.3%	84.1%	(Chou and Maggiora 1998)
Component coupled geometric classification	Amino Acid Composition	20	4 classes (SCOP)	359	Unknown, but homologous	Yes	94.4%	84.7%	(Markowetz, Edler et al. 2003) (Wei-Shu Bu 1999)
	Auto-correlation functions	65	4 classes (SCOP)	359	Unknown, but homologous	Yes	96.7%	90.5%	
Bayes classification	Amino Acid Composition	20	4 classes	131	60%	No	99.2%	42.7%	(Nakashima, Nishikawa et al. 1986) (Wang and Yuan 2000)

		20	4 classes	120	Unknown	No	100%	53.3%	(Chou 1995)
		20	4 classes SCOP	1189	40%	Yes	63.8%	53.8%	(Wang and Yuan 2000)
		20		675	30%	Yes	66.7%	48.0%	
Discriminant analysis	Amino Acid Composition and polypeptide Composition	60	4 classes SCOP	1054	40%	Yes	91.7%	75.2%	(Luo, Feng et al. 2002)
	Amino Acid Composition	20	4 classes SCOP	1054	40%	Yes	66.2%	55.8%	
Information discrepancy based classification	Polypeptides	200	4 classes SCOP	359	Unknown, but homologous	Yes	-	95.8%	(Chou and Maggiora 1998) (Jin, Fang et al. 2003)
		200		1401	30%	Yes	-	75.0%	(Jin, Fang et al. 2003)
Support vector machines	AA composition Vector	20	4 classes (SCOP)	359	Unknown, but homologous	Yes	93.0%	95.2%	(Chou and Maggiora 1998) (Cai, Liu et al. 2001)
Logistic regression	-	66	4 classes PDB	25PDB	25%	Yes	-	57.1%	(Kurgan and Homaeian 2006)

Specific tri-peptides	-	1000	4 classes PDB	25PDB	25%	Yes	-	58.6%	(Costantini and Facchiano 2008)
StackingC ensemble	-	34	4 classes PDB	25PDB	25%	Yes	-	59.9%	(Kedarisetti, Kurgan et al. 2006)
SVM (1st order polyn. kernel)	-	58	4 classes PDB	25PDB	25%	Yes	-	62.7%	(Kurgan and Chen 2007)
Fisher's discriminant algorithm	-	160	4 classes PDB	25PDB	25%	Yes	-	64%	(Yang, Peng et al. 2009)
-	-		4 classes SCOP	1189	40%	Yes	-	53.8	(Wang and Yuan 2000)
Logistic regression	-	66	4 classes SCOP	1189	40%	Yes	-	53.9	(Kurgan and Homaeian 2006)
StackingC ensemble	-	34	4 classes SCOP	1189	40%	Yes	-	54.7	(Anand, Pugalenthi et al. 2008)
-	-	-	4 classes SCOP	1189	40%	Yes	-	56.9	(Zhang, Ding et al. 2008)
-	-	-	4 classes SCOP	1189	40%	Yes	-	58.9	(Kedarisetti, Kurgan et al. 2006)

Fisher's discriminant algorithm	-	16	4 classes SCOP	1189	40%	Yes	-	65.2	(Yang, Peng et al. 2009)
-	-	-	4 classes SCOP	1189	40%	Yes	-	67.6	(Ke Chen 2008)
Specific tri-peptides	-	1000	4 classes SCOP	1189	40%	Yes	-	59.9	(Costantini and Facchiano 2008)

2.4 Conclusions

The range of topics discussed in this chapter has been necessary to introduce the work presented in the thesis, which is centred on the use of sequence-driven-features, for the prediction of structural classes. Proteins are biological compounds that are arranged in large chains of amino acids, connected by peptide bonds. These chains can vary in length, depending on the biochemical property of the protein. Genes are encoded by DNA, which is then translated into amino acid composition to represent a proteins primary structure. A protein function varies from serving antibodies for the immune response to structural proteins such as collagen, which builds body tissue. Proteins have several phases of structures that help define and build them; the first primary structure is the proteins amino acid composition sequence, which is one of the most abundant protein information available that is used to help deduce the secondary and tertiary protein structure information and many other protein properties with the aid of the correct sequence-driven-feature information.

Table 2-8 highlights the current range of classification rates, however, some of the drawbacks to these are the selection of poor quality datasets such as ones with unknown homology and small sample sizes. Classfification rates range between 43% - 96% and is due to either having high homology datasets (higher end of the % range) or sample size (lower end of the % range). Results could be read as unreliable compared to using common and high quality datasets such as 25PDB and 1189 where more robust set of results are achived 54% - 64%, 54% - 68%, respectively. These are lower but the robost and reliable results come from datasets that have large sample sizes and average homology rates are 25% and 40%, respectively. Eisenhaber (1996) et al stated the factors that influence the prediction of the structural classes of proteins are dataset size, structural class definitions and the selection of test procedures are important to consider reliable and robust predictive accuracies, which they estimated should achieve 60% (Eisenhaber, Frömmel et al. 1996; Eisenhaber, Imperiale et al. 1996). Feature sizes is also an important area to investigate as the larger the feature space the more computationally expensive it becomes to analyse with such larger feature space however it is shown in Table 2-8 that features size 1000 has a similar accuracy levels to features sizes between 20 and 100.

With all this in mind chapter 3 presents the analysis of existing materials and methods to classify protein structural classes and selects the most representative sets of materials and methods and puts forward selection of tools that is used to further explore the classification issue within this thesis.

Chapter 3 - Materials and Methods

3.1 Introduction

This chapter presents detailed information on the range of predictive assessments and methods used in the wider field of protein structural class prediction and the selected materials and methods used throughout this thesis.

3.2 Datasets

A dataset is a group of data organised in a two dimensional space, each row in a dataset represents a protein sample and each column or a group of columns represents a sequence driven feature or a feature group. Each value in a dataset is an expression of the feature for the given protein sequence. For classification for protein structural classes' datasets with large number of protein samples and low in homology are ideally suited for the robust classification of the predictability of sequence-driven features (Kurgan and Homaeian 2006).

Past studies contain a large range of classification accuracies ranging from 50% using low homology datasets to 95% using high homology datasets (Kurgan and Homaeian 2006). The higher results are based on methods that are often tested on small datasets represented by high sequence homology levels of over 50% and small sample size, which is shown and evident in Table 2-8 to have a significant impact on the prediction accuracy. Thus, it is important to select correct datasets for reliable and robust results.

The results obtained by using large sample size datasets to assess the classification of sequence-driven features are more robust and reliable than using small sample sizes. In addition to using large sample size datasets, they also need to be within a certain homology range. Datasets with sequence homology ranging between 25% - 40% tend to give more reliable and robust set of results. Where homology level is lower than 25% the effects are under training the classification algorithm, which results in lower classification accuracies (Rost 1999; Yang, Peng et al. 2010). On the other hand, where homology level is higher than 40% the effects are the classification algorithm is testing protein samples that have been used to train the classification algorithm.

Commonly used protein structural datasets that have previously been used in past studies are listed in Table 3-1. As shown most of the datasets used have either small sample sizes and/or high or unknown homology levels. These datasets are suitable for classification but not for deriving reliable and robust set of results. In order to address these issues, the selection of large sample size and low homology datasets are needed for reliable classification. Large sample size datasets are where the size is in the many thousands (>5000) which will allow for rigorous testing as it covers a larger portion of known sequences. Low homology levels between 25%-40% gives confidence that any results derived from a classification analysis has not been trained using protein sequence samples that are highly similar to one another (Dubchak, Muchnik et al. 1999; Yang, Peng et al. 2010).

In this thesis, four datasets are used to assess the classification of sequence-driven features. All four datasets are based on known structure records from the PDB that have had their structural classes manually determined by SCOP. Two of these datasets are 25PDB and 1189 as they have been used as benched datasets in previous studies (Wang and Yuan 2000; Kedarisetti, Kurgan et al. 2006; Kurgan and Homaeian 2006; Kurgan and Chen 2007; Anand, Pugalenthi et al. 2008; Costantini and Facchiano 2008; Ke Chen 2008; Kurgan, Cios et al. 2008; Zhang, Ding et al. 2008; Yang, Peng et al. 2009; Yang, Peng et al. 2010). The 25PDB dataset is selected based on the 25% PDBSELECT list using the PDB release as of February 2005 (Hobohm and Sander 1994). PDBSELECT is selection of a representative set of PDB chains (Hobohm and Sander 1994). On average the homology of all sequences is 25 % (Kurgan and Homaeian 2006). The 1189 dataset contains on average 40% homology and is constructed using the ASTRAL SCOP Genetic Domain Sequences version 1.67 database as of February 2005 (Wang and Yuan 2000).

Table 3-1 Datasets commonly used for the prediction of protein structural classes

No. of protein samples	Homology Level	Reference
260	Unknown	(Eisenhaber, Frömmel et al. 1996; Wang and Yuan 2000)
471	Unknown	(Eisenhaber, Frömmel et al. 1996; Wang and Yuan 2000)
359	Unknown but homologous	(Chou and Maggiora 1998)
131	Unknown	(Nakashima, Nishikawa et al. 1986)
120	Unknown	(Chou 1995)
1189	40%	(Wang and Yuan 2000)
675	30%	(Wang and Yuan 2000)
1054	40%	(Luo, Feng et al. 2002)
1401	30%	(Jin, Fang et al. 2003)
1601	Unknown but homologous	(Chou and Maggiora 1998)
2230	20%	(Chou and Cai 2004)
1673	25%	(Hobohm and Sander 1994)

Two additional datasets are constructed using sequences available from the ASTRAL SCOP Genetic Domain Sequences version 1.71 as of August 2008. These datasets are named Astral25 and Astral40, which have 25% and 40% homology levels, respectively. These datasets were constructed to provide the largest and most current dataset of protein samples used in the prediction of protein structural classes. Table 3-2 shows the dataset sizes.

Table 3-2 Dataset size

	No. of proteins					
Dataset	**all-α**	**all-β**	**α/β**	**α+β**	**Total**	**Reference**
25PDB	443	443	346	441	**1673**	(Hobohm and Sander 1994)
1189	223	294	334	241	**1092**	(Wang and Yuan 2000)
Astral25	1134	1273	1475	1379	**5261**	(Chandonia, Hon et al. 2004)
Astral40	1449	1730	2066	1850	**7095**	(Chandonia, Hon et al. 2004)

A custom program was developed that read PDB IDs to obtain the amino acid sequence, chain or portion of a sequence from the PDB database based on PDB ID's listed in Appendix A for two datasets 25PDB and 1189 (Kurgan and Homaeian 2006). Obtaining the Astral datasets sequences are available using its online database tool available at http://astral.berkeley.edu/ (Chandonia, Hon et al. 2004).

3.3 Dataset filtering

Dataset filtering is the process of preparing the datasets for analysis, this involved removing samples with missing values or replacing these missing values and/or pruning the dataset ready for a certain study.

One of the main sequence driven feature groups, namely PseAAC, require protein sequences with a minimum of 31 amino acid residue. This requirement is in place because, within the formula for PseAAC, the parameter λ was kept at 30. The λ is the highest tier correlation factor needed to calculate the sequence order effect (Shen and Chou 2008). The four datasets selected were therefore analysed, and the few sequences with fewer than 31 amino acid residues, were removed from the respective dataset. The final set then consists of 1085, 1668, 5257 and 7089 proteins as detailed in Table 3-3. Number of sequences with fewer than 31 amino acid residues removed from each dataset can be found in and Table 3-4 respectively.

Table 3-3 Revised datasets no. of proteins after removing sequences under 31aa from each class/dataset

	Datasets			
Structural Class	**25PDB**	**1189**	**Astral25**	**Astral40**
all-α	223	442	1132	1446
all-β	292	441	1272	1728
α/β	330	344	1474	2065
α+β	240	441	1379	1850
Total	1085	1668	5257	7089

Table 3-4 Number of sequences under 31aa removed from each class/dataset

	Datasets			
Structural Class	**25PDB**	**1189**	**Astral25**	**Astral40**
all-α	1	0	2	3
all-β	2	2	1	2
α/β	2	4	1	1
α+β	0	1	0	0
Total	5	7	4	6

To carry out further studies, which involve using both 25PDB/Astral25 and 1189/Astral40 simultaneously as testing and training datasets, respectively, identical proteins were removed from the Astral25 and Astral40 datasets that appear in 25PDB and 1189 dataset, respectively.

The Astral datasets contain a larger number of known protein structures, which overlaps the 25PDB and 1189 datasets. Table 3-5 and Table 3-6 contain the numbers of identical protein sequences and the revised astral dataset sizes after removing these sequences.

Table 3-5 Number of identical sequences found in Astral25 and Astral40 that appear in 25PDB and 1189 datasets

	Datasets	
Structural Class	Astral25	Astral40
all-α	196	98
all-β	171	114
α/β	117	138
α+β	182	113
Total	666	463

Table 3-6 Revised datasets no. of proteins after removing duplicated sequences found in Astral25 and Astral40 that appear in 25PDB and 1189 each class/dataset

	Datasets	
Structural Class	Astral25	Astral40
all-α	936	1348
all-β	1101	1614
α/β	1357	1927
α+β	1197	1737
Total	4591	6626

3.4 Classification algorithms

Classification algorithms in bioinformatics are used to classify which set(s) of categories a unknown observation from a test dataset belongs to based on prior set of training dataset of observations where categories are already known that was used to train the classification algorithm (Wu, Kumar et al. 2008).

The types of classification algorithms that are available and have been used within protein structural class classification are-

- Vector decomposition (Eisenhaber, Frömmel et al. 1996; Eisenhaber, Imperiale et al. 1996)
- Geometric classification (Chou and Maggiora 1998)
- Component coupled geometric classification (Wei-Shu Bu 1999)
- Bayes classification (Wang and Yuan 2000)
- Discriminant analysis (Luo, Feng et al. 2002)

- Information discrepancy based classification (Jin, Fang et al. 2003)
- Intimate sorting classification (Chou and Cai 2004)

There are advantages and disadvantages of each type of classification algorithm (Kurgan and Homaeian 2006). However, the most widely used classification algorithms used for the classification protein structural classes are k-nearest neighbour (KNN) classifier (Grassmann, Reczko et al. 1999; Zhang, Wang et al. 2005; Chen, Chen et al. 2008; Li, Lin et al. 2008; Melvin, Weston et al. 2008; Zhang, Ding et al. 2008; Liu, Zheng et al. 2009; Hernández-Rodríguez, Martínez-Trinidad et al. 2010) and support vector machine (SVM) ((Cai, Liu et al. 2001; Ding and Dubchak 2001; Cai, Liu et al. 2002; Cai, Liu et al. 2003; Markowetz, Edler et al. 2003; Isik, Yanikoglu et al. 2004; Chen, Zhou et al. 2006; Zhang and Ding 2007; Anand, Pugalenthi et al. 2008; Chen, Chen et al. 2008; Melvin, Weston et al. 2008; Jian-Ding, San-Hua et al. 2009; Wang, Wang et al. 2011). Both KNN and SVM are listed in the top 10 algorithms in data mining for classification (Wu, Kumar et al. 2008).

3.4.1 K-nearest neighbour classifier

KNN is the simplest of all machine-learning algorithms and is used in a number of different bioinformatics fields. Application of KNN in bioinformatics goes into gene expression for clinical prediction (Parry, Jones et al. 2010), protein-protein interaction and tertiary structure prediction and identifying biomarkers with feature selection (Li, Umbach et al. 2004). The KNN classifier works by assigning the query test protein *q* a class that is decided by its nearest k neighbours majority class member. The nearest k neighbours are the training dataset of sample proteins. An illustration of KNN is shown in Figure 3-1 with further explanation.

The method used to decide the nearest k neighbour majority class label of a test protein *q* is the Euclidean distance, which is defined in Eq 3-1.

$$D(q_1, p_n) = \sqrt{(q_1 - p_1)^2 + (q_1 - p_2)^2 + \cdots + (q_1 - p_i)^2 + (q_1 - p_{i+1})^2} \quad \textbf{Eq 3-1}$$

Where D is the Euclidean distance between query test protein *q* and training proteins *p* and *n=1…i* the number of training proteins *p*. Euclidean distance method calculates the straight line distance between two data points at a time and add the next training data points i.e. between query test protein *q* and nearest training data points proteins *p* (Metfessel, Saurugger et al. 1993).

The selection of *k* (nearest k neighbours) is very important and user dependent by the type of classification issue at hand. Usually *k* equals to the number of known classes the test protein *q* could potentially categories into (Zhang, Wang et al. 2005). Small value of *k* may lead to noisy data will have a higher influence on the results and a large value of *k* makes it computationally expensive to run the classification algorithm. What has not been investigated before is a range of *k* values as studies typically set *k* to a fixed value (Hotta, Kiyasu et al. 2004; Zhang, Wang et al. 2005). The analyses to be carried within this thesis will look at the effects of where *k* is set 1 to 11 this will allow investigation into the effects of a wide range of *k* values which has not been investigated before, in particular where *k* is greater than the number of structural classes (*k*>4). The maximum *k* value (*k*=11) was chosen since it covers a larger than usual selection *k* range but not too high where it becomes computationally expensive to run analyses as more time and resources are required to complete classification analysis. Odd *k*'s will avoid ties when selecting nearest *k* neighbours of testing proteins *p*, whereas, even numbered *k*'s will result in a random selection of the nearest *k* neighbour. More information on nearest *k* neighbours selection is shown in Figure 3-1.

Figure 3-1 illustrates how the KNN classifier depends on the selection of the *K* parameter in classifying the class of test protein *q*. Training proteins *p* belong to four different classes, class 1 is represented by red circles, class 2 are blue circles, class 3 are green circles and class 4 are brown circles. When *k* = 1 the test protein sample *q* is classified into class 2 as its nearest training protein sample is class 2. Where *k*=3, the test protein *q* is classified into class 3 because two of its three nearest training proteins *p* belong to class 3. When *k*=5 the test protein *q* is in a tie the query test protein sample *q* belongs to either class 2 or 3 because there are two of each of 3 nearest proteins that fall into class 2 and class 3 KNN will randomly assign a class to the query protein. Finally when *k*=11 the nearest protein belongs to class 2.

Multiple k-nearest neighbour (MKNN) classifiers extends KNN by combining and analysing the strongest (in terms of classification accuracies) set of *K's* to achieve the same or better result than using a single *K* in KNN will result in the overall accuracy of the combined set of strongest *k's*. Figure 3-2 illustrates how MKNN stakes into consideration the strongest resulted *ks*. Each analysis will produce 2047 (=2^11-1) classification accuracies, with each analysis carried out using the MKNN classifier the results obtained will cover each *k*=1 to 11.

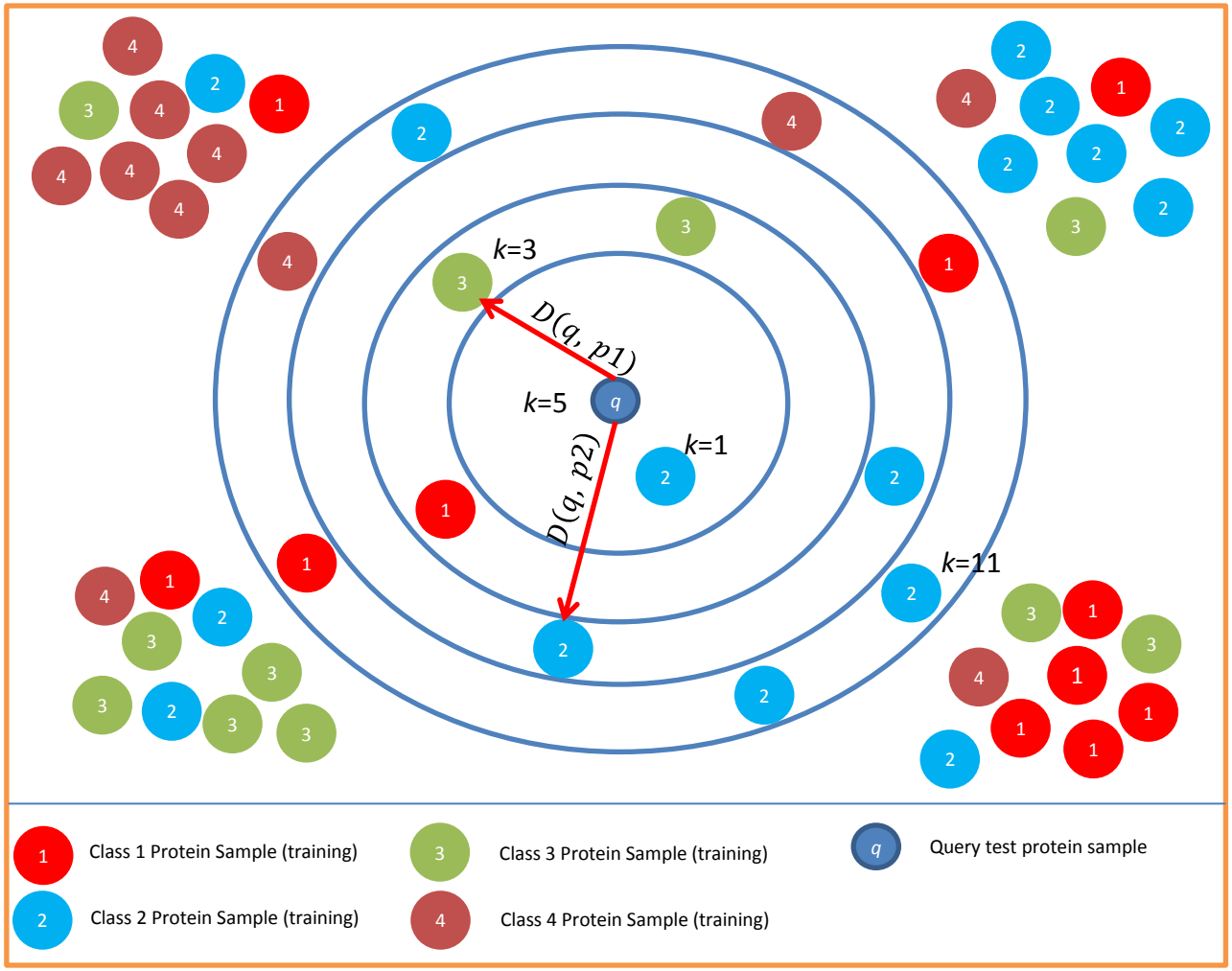


Figure 3-1 k-nearest neighbour classification

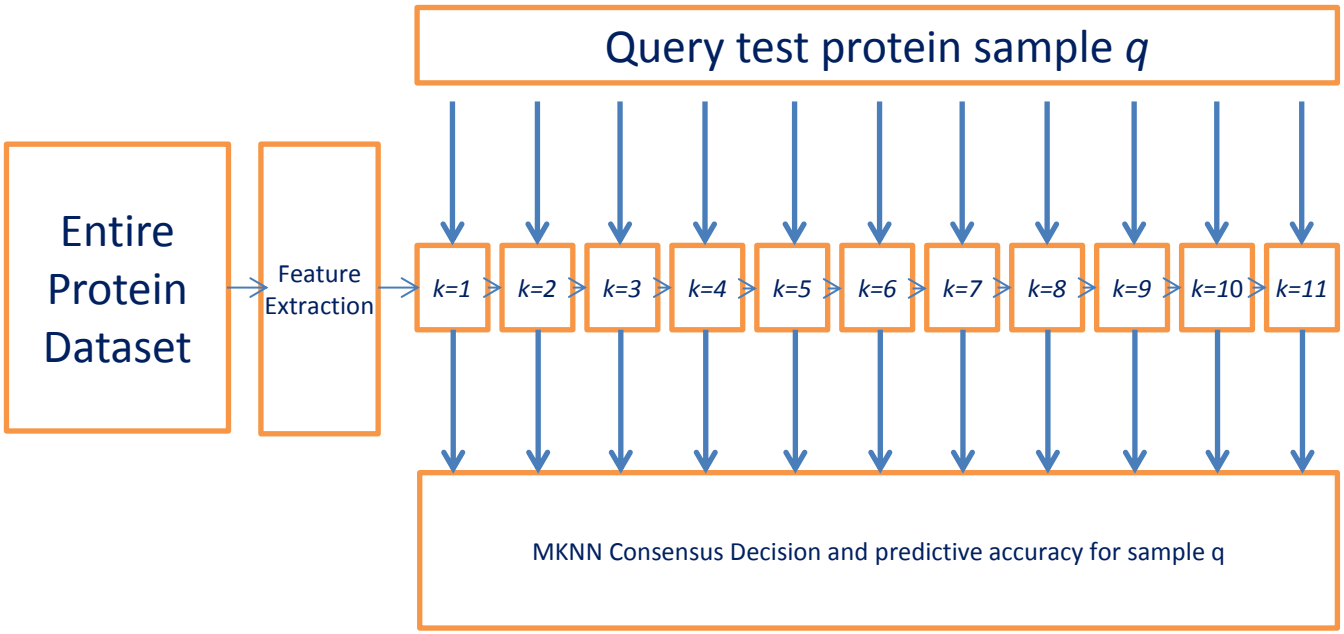


Figure 3-2 multiple k-nearest neighbour

Table 3-7 Example output of MKNN voting

K	Class 1 all-α	Class 2 all-β	Class 3 α/β	Class 4 α+β
1	1	0	0	0
2	1	0	0	0
3	0	1	0	0
4	1	0	0	0
5	1	0	1	0
6	1	0	0	1
7	0	0	0	0
8	1	1	0	0
9	1	0	0	0
10	0	0	1	0
11	1	0	0	0
Total	**8**	**2**	**2**	**1**

Table 3-7 is an example output of the MKNN voting result. Each *k* assigns a class label to the query protein. After 11 *k's* the result is tallied up and the class with largest tally wins the *mknn* result and is assigned that class label.

3.4.2 Support Vector Machine

Support vector machine (SVM) (Cortes and Vapnik 1995) is a relatively new algorithm that has been gaining popularity in a wide range of studies and designed to handle high-dimensional data (Chang and Lin 2001; Cai, Liu et al. 2003; Isik, Yanikoglu et al. 2004; Ma, Wu et al. 2009) (Cai, Liu et al. 2002; Chen, Tian et al. 2006; Kedarisetti, Kurgan et al. 2006; Zhang and Ding 2007; Anand, Pugalenthi et al. 2008; Chen, Chen et al. 2008; Chen, Zhang et al. 2008; Costantini and Facchiano 2008). Bioinformatics application in protein structural classes studies have used SVM (Cai, Liu et al. 2003; Markowetz, Edler et al. 2003; Chen, Zhou et al. 2006; Ding, Zhang et al. 2007; Zhang, Wei et al. 2007; Anand, Pugalenthi et al. 2008; Wang, Wang et al. 2011). SVM maps data points into a high dimensional feature space and tries to find separating hyperplane with the maximal margin to distinguish different classes. First step of SVM is to map out the input vectors (training protein samples) into a feature space, linearly or non-linearly, which is appropriate to the selection of the SVM kernel function. Within the feature space the first step is to seek an optimised division, i.e. construct a hyperplane which separates at least two this will separate the predicted data points in its respected predicted class (Markowetz, Edler et al. 2003). SVM training always tries to find an optimised solution

and tries to avoid over-fitting the data, so it has the potential ability to deal with a large number of features (Cortes and Vapnik 1995; Cai, Liu et al. 2001; Chang and Lin 2001). Figure 3-4 is an example of a linear SVM classifier, i.e., a classifier that separates a set of data points into their respective groups (red and blue in this case) with a line. Most classifications are not linearly solvable as they have more than two classes of categories to distinguish; these classification tasks are more complex and require a different type of SVM approach to data separation, a typical example is shown in Figure 3-3. The original optimal hyperplane algorithm proposed was a linear classifier by Vapnik in 1963. In 1992, Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik suggested a new way to create a nonlinear classifier by applying a kernel function (Aizerman, Braverman et al. 1964) to maximize the margins of the hyperplanes (Boser, Guyon et al. 1992). Kernels are the functions that determine the different mapping of vectors in the feature space (Chang and Lin 2001). SVM is parametric classifier and is very sensitive to selection of parametric values that kernel function require. To find the optimum set of parameters for the prediction problem in hand a grid search should be used to find the best *C,* where *C* is the penalty parameter. The penalty parameter balances the trade-off between training accuracy and generalisation of the classifier, the larger the value *C* the more the error is penalised, but too large it will try to find the best fit for all the training data points (Anand, Pugalenthi et al. 2008). Therefore *C* parameter should be chosen with care to avoid over fitting the predictive model (Yousef, Ketany et al. 2009). The resulting SVM nonlinear algorithm is similar to linear SVM, except that every dot product is replaced by a nonlinear kernel function. This will allow the SVM algorithm to fit the maximum-margin hyperplane in a transformed feature space. The transformation may be nonlinear and the transformed space is high dimensional; thus though the SVM classifier is a hyperplane in the high-dimensional feature space, it may be nonlinear in the original input space (Hmeidi, Hawashin et al. 2008; Ma, Wu et al. 2009). An example of a more typical classification problem is shown in Figure 3-3 which is the non-linear SVM classification. Compared to Figure 3-4 separating the red and blue data points will require a curve. Dividing data points using a curve is more difficult than using a line (Cortes and Vapnik 1995).

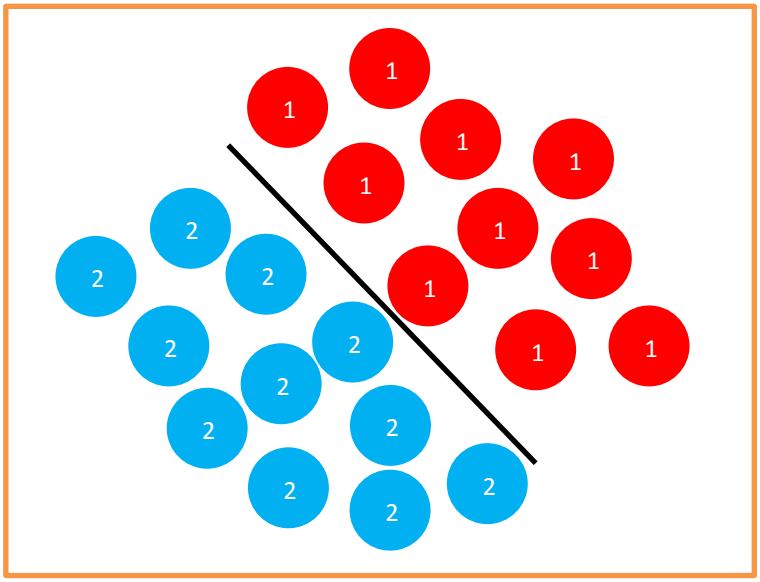


Figure 3-4 Linear SVM

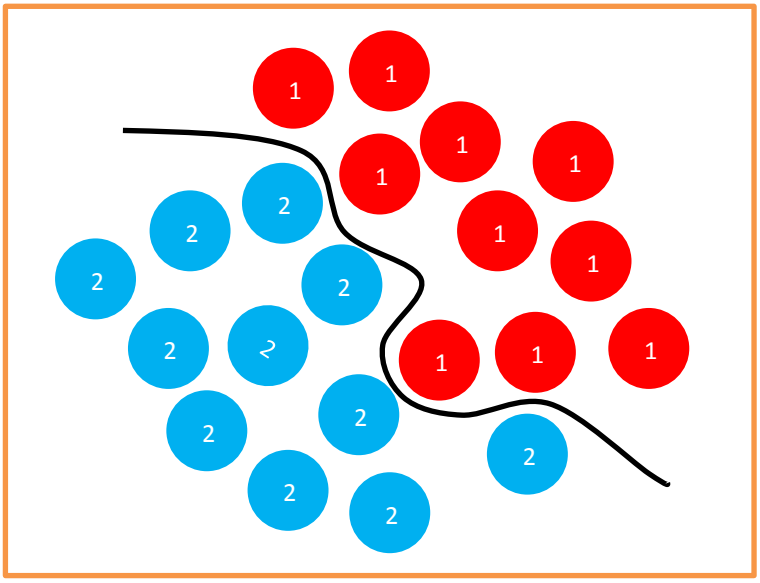


Figure 3-3 SVM non-linear classifier

3.4.3 Differences between KNN and SVM

The differences between KNN and SVM algorithms, is that KNN is a supervised machine learning algorithm and non-parametric that reads a set of labelled training data and then uses them to classify an unlabelled testing sample. In order to classify a test protein sample KNN calculates the Euclidean distances between the testing protein and all the training proteins. Then the K nearest training proteins to the test protein are used to classify the structural class of the testing protein (Zhang, Wang et al. 2005). SVM is also a supervised machine-learning algorithm and parametric, but the result of SVM is sensitive to the selection of the *C* parameter. SVM tries to finds the maximum margin hyperplane by separating distinguishable classes in data, if the data points are not linearly separable in the feature space, as is often the case, they can be projected into a higher dimensional space by means of a kernel function (Cortes and Vapnik 1995). Analysis results using SVM did not produce better accuracies compared to MKNN as the selection of parameters were very sensitive across each dataset, the main prediction algorithm used throughout all the analyses was the multiple-k-nearest neighbour (MKNN) classifier.

3.4.4 Classification performance

Evaluating the performance of classifiers is called class performance that outputs the classification accuracy, which is defined as correctly classified proteins / classified proteins as shown in Eq 3-2.

$$\text{accuracy} = \frac{\text{No. of correctly classified proteins}}{\text{No. classified proteins}} \qquad \textbf{Eq 3-2}$$

Table 3-8 describe how the numbers of correctly classified samples are produced via a confusion matrix to produce classification accuracies. Where a, b, c and d is shown represents the number of correctly classified proteins for each of the structural classes of proteins all-α, all-β, α/β, α+β respectively. Where ab is shown it is the number of incorrectly classified all-α proteins as all-β proteins, similarly, predictions where cd is shown are α/β incorrectly classified as α+βall-α protein is predicted as all-β protein, ba is the number of incorrect predictions where all-β protein is predicted as all- α protein, etc.

Table 3-8 Confusion matrix for the protein structural class prediction

Actual Structural Class	Classified Structural Class			
	All-α	**All-β**	**α/β**	**α+β**
All-α	**a**	ab	ac	ad
All-β	ba	**b**	bc	bd
α/β	ca	cb	**c**	cd
α+β	da	db	dc	**d**

An example of an analysis predictive result is given in Table 3-9; it shows the accuracies of correctly classified proteins for each structural class and then the overall accuracy for each *k*. The individual class accuracy is calculated by counting how many proteins from each structural class are correctly classified and divided by the total number of proteins in the respective class e.g. the accuracy for all-α = 59.19% (where *k*=1 highlighted bold in Table 3-9). The result is the calculation of 132/223, where 132 is the number of correctly classified all-α divided by the total number proteins in the all-α structural class.

Table 3-9 Example set of an analysis result

K	All-α	All-β	α/β	α+β	Overall
1	**59.19%**	63.36%	66.97%	56.25%	62.03%
2	59.19%	63.36%	66.97%	56.25%	62.03%
3	49.78%	56.51%	62.42%	47.92%	55.02%
4	43.05%	54.79%	62.73%	43.33%	52.26%
5	43.95%	53.42%	63.33%	46.67%	53.00%
6	44.39%	53.77%	61.21%	40.00%	51.06%
7	39.46%	50.68%	63.03%	43.75%	50.60%
8	38.57%	51.37%	64.24%	41.67%	50.51%
9	36.77%	51.71%	64.24%	40.42%	49.95%
10	38.12%	50.34%	68.48%	40.00%	51.06%
11	35.43%	51.71%	67.27%	36.25%	49.68%
12	59.19%	63.36%	66.97%	56.25%	62.03%

3.4.5 Test procedures

The data partitioning of datasets is done using three different test procedures (a) n-fold cross-validation, (b) leave-one-out and (c) independent-sets. These test procedures are interchangeably called cross validation methods in the literature, in this thesis we refer to them as test procedures. The different test procedures each alter the size of testing and training subsets of the whole dataset and how the predictive model accesses them, subsets of the data are held out to be used as testing sets; the predictive model is then fitted to the training set and used to predict the testing subset. It is a crucial aspect to any classification

method that the training and testing subsets of the dataset must be independent of each to avoid sub sampling which leads to unreliable results. The primary reason for using three different test procedures is to obtain a consensus set of results for each analysis as it could be shown that one test procedure maybe better in one type of analysis than another.

3.4.5.1 N-fold cross-validation

A dataset is partitioned into N folds where approximately No. of proteins / N percentage of the protein samples used per fold, where N is the number proteins in the dataset. One fold is testing data (Blue Square) in Figure 3-5; the remaining N-1 folds are used for training. This partitioning repeats until each fold is tested and the results are averaged across all N folds. The advantage of N-fold cross validation is that all the samples in the dataset are eventually used for both training and testing (Metfessel, Saurugger et al. 1993; Grassmann, Reczko et al. 1999; Kurgan and Homaeian 2006). The disadvantage of this test procedure is that the training has to be completed N number times. For the sake of consistency, the arrangement of protein samples in each fold should be kept the same throughout all the analyses. Figure 3-5 describes the N-fold cross validation test procedure where the number of folds equal to *N = 10* which is used in the PhD thesis.

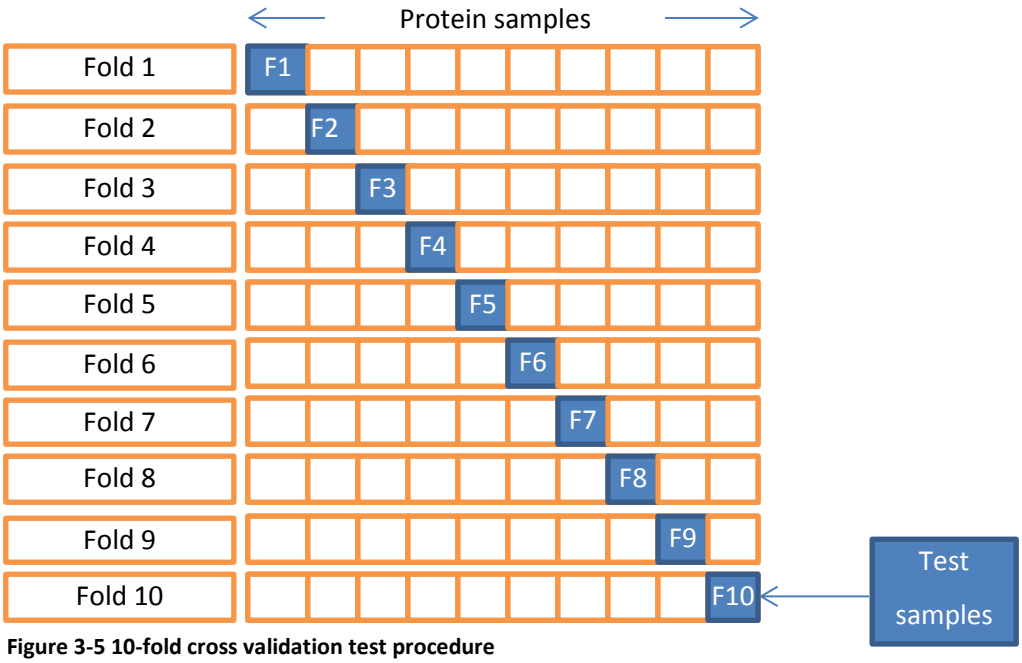


Figure 3-5 10-fold cross validation test procedure

3.4.5.2 Leave-one-out

Leave-one-out is a similar process to N-fold, but *N* = total number of proteins. During the leave-one-out process, a single protein will move from training dataset to testing dataset, once the test sample is analysed the single protein is then moved back to training dataset. Testing dataset only contains one protein to test and the training dataset is the remaining set of proteins N-1. This is a computationally demanding and resourceful technique because of process of testing each protein sample one at a time, this more particular for larger datasets. The advantage is that it applies a thorough testing to each protein sample and no random sub sampling (Boser, Guyon et al. 1992; Kertész-Farkas, Dhir et al. 2007). Figure 3-6 illustrates leave-one-out test procedure.

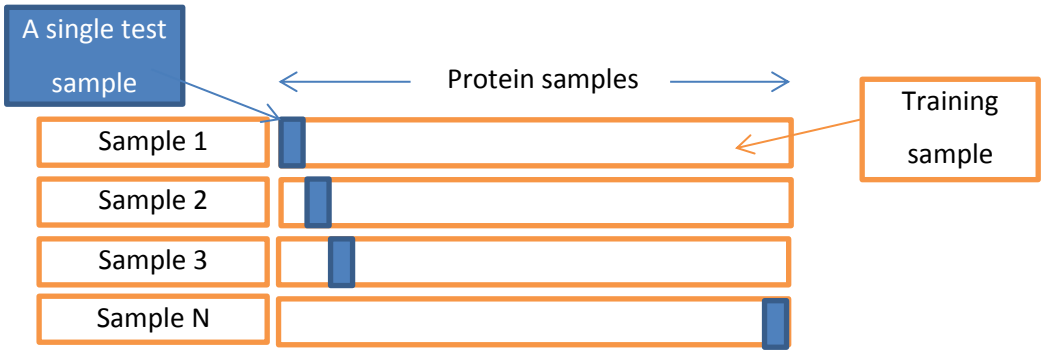


Figure 3-6 Leave-one-out cross validation test procedure

3.4.5.3 Independent-sets

Independent test procedure uses two separate datasets for testing and training. These datasets contain unique protein sequences i.e. there are no two identical protein sequences between testing and training datasets (Mizianty and Kurgan 2009). The construction of independent datasets involved comparing each protein sequence from the training dataset with each protein sequence from the testing dataset, this involved looping between two matrices (i.e., testing and training datasets) to identify duplicate sequences. If the comparison of sequences equalled to one, it means a duplicate sequence is found and thus removed from either of the astral dataset. Revised Astral dataset sizes are shown in in Table 3-10.

Table 3-10 Independent training and testing dataset combinations

Training Dataset	Dataset Size	Testing Dataset	Dataset Size
25PDB	1668	Astral25	4591
1189	1085	Astral40	6626
Astral25	4591	25PDB	1668
Astral40	6626	1189	1085

3.5 Hierarchical clustering

Clustering is the process of grouping a set of objects in a way that objects in the same group (called cluster) are more similar (in some context or another) to each other than to those in other clusters. It is an important task in data mining, machine learning, pattern recognition for bioinformatics.

Hierarchical clustering group data over a variety of measures by creating a cluster tree or dendrogram (Davies, Secker et al. 2007). The cluster tree is not a single set of clusters, but a multilevel hierarchical, where clusters at one level are joined at clusters at the next level. Hierarchical clustering technique associates patterns found in data via a linkage line by measuring the similarity between data points, the similarity method is based on the Euclidean distance function.

To perform cluster analysis on a data set, the following steps should be followed (1) find the similarity or dissimilarity between every pair of data point in the data set, the similarity or dissimilarity is based on the distance between data points. (2) Group the data points into a binary, hierarchical cluster tree, and this links pairs of data points that have close distance values using the linkage function. The linkage function uses the distances information to determine the proximity of the data points to each other. (3) Determine where to cut the

hierarchical tree into clusters (cut off point), prune branches off the bottom of the hierarchical tree and assign all the data points below each cut to a single cluster. The general idea is to merge data points into similar groups until a threshold value is met.

There are two approaches to hierarchical clustering (1) from bottom up which combines small clusters into big ones and (2) from the top down which breaks up a large cluster of data points into smaller ones (Goldstein, Zucko et al. 2009). These are also known as agglomerative and divisive hierarchical clustering. Three types of agglomerative hierarchical clustering are adopted, which are:

- Single linkage clustering that uses the minimum distance between objects in clusters.

$$D(r,s) = \min(dist(x_{ri}, x_{sj})) \quad \textbf{Eq 3-3}$$

- Complete linkage clustering that uses the maximum distance between objects in clusters.

$$D(r,s) = \max(dist(x_{ri}, x_{sj})) \quad \textbf{Eq 3-4}$$

- Average linkage clustering that uses the average distance between all pairs of objects in clusters.

$$D(r,s) = \frac{1}{n_r\, n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri}, x_{sj}) \quad \textbf{Eq 3-5}$$

where $i \in (1\ldots, n_r)$, $j \in (1, \ldots, n_s)$, n_r and s_n are the number of objects in cluster r and s, respectively. x_{ri} the ith object in cluster r and x_{sj} is the jth object in cluster s.

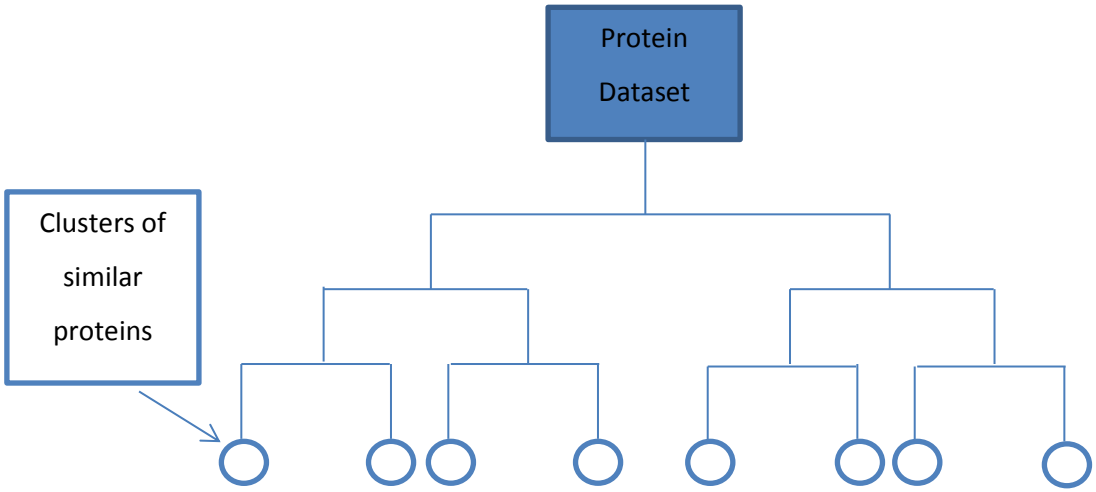


Figure 3-7 A visual example of agglomerative hierarchical (from the top down) clustering

3.5.1 Bioinformatics application of hierarchical clustering

Hierarchical clustering is a powerful data mining method to exploring proteomic data. It enables proteomic data to be grouped blindly according to their expression values without any prior knowledge of the biology behind the data (Meunier, Dumas et al. 2007). One of the main hierarchical clustering bioinformatics applications is the exploration and defining of natural groups within the data. Arnau et al. (2003) implemented hierarchical clustering to protein-to-protein interaction data, the original data contains many sorts of different interactions some of these different interactions are quite similar where it may be difficult to group similar data points, hierarchical clustering aided the separation of data points into clusters of similarities. Another application of hierarchical clustering is the clustering of highly homologous sequences to reduce the size of large protein databases (Li, Jaroszewski et al. 2001) and clustering of physicochemical and biochemical properties of amino acids (Saha, Maulik et al. 2011).

3.6 Principal Component Analysis

Principal component analysis (PCA) is a feature reduction method that can be used to reduce a large number of features into a small set of artificial features that are called principal components; these principal components can represent most of the information (variance) of original set of features in as few as possible. All the principal components are orthogonal (uncorrelated) to each other so there is no redundant information in the new set of artificial features (Wang 2003; Du, Jiang et al. 2006).

The first principal component is a single axis in the data space [add pic]. When data points from the feature set are projected on to that axis, the resulting values forms a new feature set

and the variance of this new artificial features set is the maximum among all possible choices for the first axis. The second principal component is another axis in the data space, perpendicular to the first principal component. Projecting data points from the feature set on to this axis generates another new artificial features set. The variance of this new feature set is the maximum among all possible choices for the second axis. This will continue until all the original features have been transformed into principal components. The full set of principal components is as large as the original set of features.

Each principal component represents a certain variance of the original set of variables, the first principal component usually accounts for the largest proportion of the original data variance and additional principal components account for the remaining amount, although a large number of principal components can be extracted to obtain 100% variance. However, it is common for the sum of the variances of the first few principal components to cover majority of the total variance of the original features (Wang 2003; Du, Jiang et al. 2006). The component analysis produces components in descending order of importance – (i.e., the first component explains the maximum amount of data variation and the last component the minimum amount of data variation). is an example of a set of PCA result, the original feature space contained 2 variables (of 20 descriptor values each), after PCA is applied the original feature space is no longer visible as PCA converted the 2 possibly correlated feature sets of into a set of 2 uncorrelated features. The number of principal components is equal to the number of original features, which is 2. The first principal component has the largest (in this example all) possible variance and the second principal component contains the least amount (in this can zero), the principal components are orthogonal to each other. The example shown resulted in that principal component 1 contains 100% of the variability of the original feature space. In practical terms instead of using a feature space of two, the data can be reduced to a feature space of one (i.e. principal components) which will cover just under 100% of the original data.

Table 3-11 Example of PCA components and variances that represents the data

Component N.	Component variance:	Percentage of variance
1	1	100%
2	0	0%

3.6.1 Bioinformatics application of PCA

In bioinformatics the high dimensionality of datasets is a field wide issue (Goodman and Hunter 1999), where the protein samples are outnumbered by the number of features. PCA is widely used method in bioinformatics in areas such as genomic and proteomic studies for mean reason to reduce the size of large datasets. An example of large datasets is with gene expression data, the number of protein size is much smaller than the number of genes and the dimensionality of gene expressions needs to be reduced prior to any classification and/or clustering types of analyses. This remove noisy data without removing representative data which can reduce computational time to analyses the smaller dataset which could improve classification results, PCA has to been shown to have satisfactory performance (Ma and Dai 2011).

3.7 Conclusions

Chapter 3 presents the materials and methods that are in use within the field and the selection of them to be used within this thesis and the reasons for them. Correct selection of currently used in the field datasets 25PDB and 1189, this will allow comparable results with other studies using benched mark datasets. The construction of largest size structural classes of proteins dataset is very important, as using large sample sizes with low homology levels will give more reliable results compared to using low size datasets that have high homology levels. The largest datasets are named Astral25 and Astral40, the number denotes homology level, which each contain over 5000, and 7000 proteins, respectively, will allow the methods developed and presented in the next three chapters to be tested against. The selection of dataset homology levels is another important factor that has been considered, as previously discussed results that are obtained using homology levels between 25-40% are more reliable and consistent. Compared to levels higher than 40% result are considered unreliable as they tend to be as close as 90% accuracy levels too low will makes classification harder as some prior information is needed for the classifier to be trained against. The selection of classification algorithm is another factor to consider, as it is necessary to classifiers that will be independent and unbiased towards any datasets or methods; however, the aim of the thesis is not to develop a classifier but methods that can be used with any classification algorithm. m*knn* is a fast and simple non-parametric and supervised classifier to use, which uses a voting method to find the overall strongest k models for a better result than using a single *k*. Another area that was weak within the literature was the investigation into the use and impact of

multiple test procedures, it was noted that leave-one-out is the most widely used test procedure, which was shown to be a reliable method but was very computationally demanding. This thesis will consider the use of n-fold and independent-sets test procedures, which uses independent datasets for training and testing subsets. Chapter 4 investigates the largest set of the most widely used sequence-driven-features for the prediction of structural classes of proteins using the selection of materials and methods mentioned above.

Chapter 4 - Analysis of existing sequence driven features

4.1 Introduction

The investigation into the role of sequence-driven features for the prediction of protein structural classes has not been investigated in the field. Chapter 4 presents the analysis into the largest set of sequence-driven features for the prediction of protein structural classes using the selection of materials and methods from chapter 3. The objective is to carry out a comprehensive and consistent investigation over the largest set of protein sequence-driven features that forms 10 feature groups with 53 further subsets of the features to evaluate and identify how well these features predict protein structural classes. The two outcomes are (i) which of these sequence-driven feature group(s) or sub-feature(s) are more suitable for the prediction of structural classes of protein, and (ii) to develop a benched mark set of results that will form the baseline to which previous published work and future work will be compared against. Chapter 4 will also look into the results of how the four different datasets and three different test procedures can affect the ranked order of features based on classification accuracies.

4.2 Sequence representation: Sequence-driven features

Sequence driven features are various structural and physicochemical properties extracted from protein amino acid sequences and appropriate selection is a crucial part in prediction of structural classes of proteins. For this study the largest set of sequence-driven features are extracted for each protein dataset 25PDB, 1189, Astral25 and Astral40 and then analysed for the prediction of structural classes of protein. Protein sequences from the datasets have varying sequence lengths to feed these sequences into the *k*nn classifier the sequences must be converted from a varying length of amino acid letters into fixed length of numerical vectors (Kyoung Kim, Bang et al. 2006).

For the prediction of structural classes of proteins the most common sequence-driven features groups that are used are AAC and PseAAC feature groups; however, there are many other feature groups available that need be investigated such as dipeptide composition, autocorrelation, composition, transition, distribution and sequence order – these feature groups are discussed later in this chapter. Utilisation of various sequence-driven features may better help characterise and discriminate proteins of different structural properties by exploring their distinguished features in compositions, correlations, transitions, distributions and sequence-order of the constituent amino acids and their structural and physicochemical properties (Li, Lin et al. 2006; Ong, Lin et al. 2007).

Chapter 4 evaluates the largest set of sequence-driven features for their effectiveness for predicting structural classes of protein by using a consistent set of materials and methods as it is of great interest to find out which set of sequence-driven features helps to improve classification accuracy. Appendix I contain the full set of sequence-driven features used in the study 10 feature groups with 53 sub features, each feature group, sub feature and descriptor value has a feature index number and its feature size included in the appendix.

4.3 Sequence driven features technical details

This section will explain the theory behind the ten feature groups and its feature subsets. Amino acid composition is the first feature group with 20 descriptor values. The second feature group is dipeptide composition with 400 descriptor values. The third, fourth and fifth feature groups are three different autocorrelation based feature groups, normalized moreau–broto autocorrelation, moran autocorrelation and geary autocorrelation. Each autocorrelation feature group has eight sub features containing in total 240 descriptor values. The sixth, seventh and eight feature group are composition, transition and distribution each containing 21, 21, 105 descriptor values, respectively. The ninth feature group composes of two sub sequence-order feature sets; the first sub feature is sequence-order-coupling number with 60 descriptor values and the second sub feature quasi-sequence-order and 100 descriptors values and the final tenth group is PseAAC. The following sub sections describe each feature group in more detail; the equations are taken from the supplementary manual from the ProFEAT website available at http://jing.cz3.nus.edu.sg/cgi-bin/prof/prof.cgi

4.3.1 Amino Acid Composition

One of the earliest methods and the most common ways to represent a protein's amino acid sequence is to compute amino acid composition of the protein. Amino acid composition calculates the fraction of each amino acid type in a protein sequence. Twenty descriptor values are computed for the 20 types of amino acids. The original proposal to use amino acid composition was developed by P.Y Chou in 1980 and later in 1982 Nishikawa et el (Nishikawa and Ooi 1982; Nakashima, Nishikawa et al. 1986) cited that the amino acid composition of a protein is correlated with structural class. Since the late 80's representing proteins by its 20 dimensions feature vector, progress in the prediction of the structural classes of proteins have been made using various methods coupled with amino acid composition (Chou 1995; Eisenhaber, Frömmel et al. 1996; Eisenhaber, Imperiale et al. 1996) (Nakashima, Nishikawa et al. 1986; Zhang and Chou 1992; Chou 1995; Chou and Zhang 1995; Eisenhaber, Frömmel et al. 1996; Eisenhaber, Imperiale et al. 1996; Bahar, Atilgan et al. 1997; Luo, Feng et al. 2002; Du, Jiang et al. 2006; Nanuwa and Seker 2008; Seker 2008; Nanuwa, Dziurla et al. 2009).

Amino acid composition calculates the fraction of each amino acid type in a sequence as defined in Eq 4-1 (Li, Lin et al. 2006).

$$f(r) = \frac{N_r}{N} \qquad \textbf{Eq 4-1}$$

where r = 1, 2, 3, . . . , 20, Nr is the number of amino acid of type r, and N is the length of the sequence. Twenty descriptor values are calculated (the 20 types of amino acids) for amino acid composition. Table 4-13 contains the top 10 results for amino acid composition analyses and where feature index 1 is present indicates the predictive accuracies obtained through the analysis of AAC. An example of amino acid composition is shown in Table 4-1. The amino acid sequence presented in Figure 2-4 belongs to the crystallization of human beta3 alcohol dehydrogenase protein (PDB ID 1HTB) is converted into amino acid composition feature. Each value under each amino acid letter is the total composition of that amino acid in the sequence, e.g. amino acid A (Alanine) represents 8.28% of the number of residues present in the sequence. The total of all values comes to 100, which represents 100% composition of the amino acids presented in the sequence.

Table 4-1 Amino acid composition example of protein 1HTB

Amino Acid	A	C	D	E	F	G	H	I	K	L
Fraction Value	8.289%	4.278%	4.545%	5.080%	4.278%	10.160%	1.872%	5.882%	8.556%	7.754%
Amino Acid	M	N	P	Q	R	S	T	V	W	Y
Fraction Value	1.872%	3.209%	5.348%	1.604%	2.673%	5.614%	6.417%	10.42%	0.534%	1.604%

4.3.2 Dipeptide Composition

Dipeptide composition calculates the fractions of pairs of amino acids i.e. it will search for all AA, AC, AD, AE etc. and then next set of amino acids CA, CC, CD, CE and then next DA, DC, DD, DE etc. defined in Eq 4-2 (Li, Lin et al. 2006).

$$fr(r,s) = \frac{N_{rs}}{N-1} \qquad \textbf{Eq 4-2}$$

where r,s = 1, 2, 3, . . . , 20, and N_{rs} is the number of dipeptides of amino acid type r and s. Four hundred descriptor values are calculated for the 20 x 20 amino acid combinations. The biological significance of dipeptide composition towards protein sequences is that it encapsulates the global information of AAC and local order of amino acids in a protein sequence, so it considers all of the adjoining pairs of amino acids whether they are identical or non-identical (Bhasin and Raghava 2004). Dipeptide composition has been used in studies such as prediction of subcellular locations of proteins and fold recognition (Reczko and Bohr 1994; Grassmann, Reczko et al. 1999). Feature index 2 in Table 4-13 (if present) indicates the dipeptide composition predictive accuracies.

A snippet example of 40 out of 400 features of dipeptide composition using the sequence presented in Figure 2-4 is shown in Table 4-2 the total value of all 400-dipeptide descriptors comes to 100, which represents 100% of the pairs of amino acids.

Table 4-2 Dipeptide composition example

AA	AC	AD	AE	AF	AG	AH	AI	AK	AL
1.340483 %	0.268097 %	0.268097 %	0%	0%	0.80429 %	0%	0%	1.340483 %	0.268097 %
AM	**AN**	**AP**	**AQ**	**AR**	**AS**	**AT**	**AV**	**AW**	**AY**
0%	0%	0.268097 %	0%	0.268097 %	0.80429 %	0.268097 %	2.144772 %	0%	0.268097 %
CA	**CC**	**CD**	**CE**	**CF**	**CG**	**CH**	**CI**	**CK**	**CL**
0.268097 %	0.268097 %	0%	0%	0%	0.80429 %	0.268097 %	0.268097 %	0.80429%	0.536193 %
CM	**CN**	**CP**	**CQ**	**CR**	**CS**	**CT**	**CV**	**CW**	**CY**
0%	0%	0%	0%	0.80429%	0%	0.26810%	0%	0%	0%

4.3.3 Autocorrelation feature groups

The autocorrelation sequence-driven feature groups describe the correlation between protein amino acid sequences in relation to their specific structural or physicochemical property, which are based on the distribution of amino acid properties along the protein amino acid sequence (Broto, Moreau et al. 1984).

1. Normalized moreau-broto autocorrelation
2. Moran autocorrelation
3. Geary autocorrelation

The difference between moran autocorrelation and moreau-borto autocorrelation is the use of the amino acid index property deviations from the averages values instead of the index property values themselves as the basis for measuring correlations. The Geary autocorrelation feature compared to others uses Square differenced of property values.

Each autocorrelation group has eight sub features and each is an amino acid index, which represent different physicochemical or biochemical properties of the amino acids. Table 4-3 list the eight subsets of autocorrelation sub features.

Table 4-3 Subset of the autocorrelation features

Subset feature No.			Subset feature	Amino acid index (refer to appendix II)	Reference
3.1	4.1	5.1	Hydrophobicity scale	58	(Cid, Bunster et al. 1992)
3.2	4.2	5.2	Average flexibility Indices	8	(Bhaskaran and Ponnuswamy 1988)
3.3	4.3	5.3	Polarizibility parameter	22	(Charton and Charton 1982)
3.4	4.4	5.4	Free energy in water	23	(Charton and Charton 1982)
3.5	4.5	5.5	Residue accessibility surface area in Tripeptide	33	(Chothia 1976)
3.6	4.6	5.6	Residue volume	9	(Bigelow 1967)
3.7	4.7	5.7	Steric parameter	21	(Charton 1981)
3.8	4.8	5.8	Relative mutability	65	(Dayhoff 1978)

The eight amino acid indices utilised in each autocorrelation feature groups are:

1. Hydrophobicity scales (feature index 3.1, 4.1, and 5.1) - Cid et al. described hydrophobicity scales as the values that define hydrophobicity of amino acid residues i.e. that is repelled from water (Cid, Bunster et al. 1992).
2. Average flexibility indices (feature index 3.2, 4.2, and 5.2) - Bhaskaran et al. described average flexibility indices as values that define symmetric/asymmetric distribution of

amino acid residues i.e. that is protein structural flexibility (Bhaskaran and Ponnuswamy 1988).

3. Polarizability parameter (feature index 3.3, 4.3, and 5.3) - Charton et al. describes the polarizability parameter as part of the structural dependence of amino acid hydrophobicity parameters (Charton and Charton 1982).
4. Free energy of solution in water (feature index 3.4, 4.4, and 5.4) - Charton et al. describes the free energy of solution in water parameter as part of the structural dependence of amino acid hydrophobicity parameters (Charton and Charton 1982).
5. Residue accessible surface area in trepeptide (feature index 3.5, 4.5, and 5.5) – Chothia described residue accessible surface area in trepeptide as values that define interactions with water is central secondary structure. Where "accessible surface area " describe the extent to which atoms on the protein surface can form contacts with water and trepeptide is a peptide (a peptide is a protein with a small amino acid sequence) consisting of three amino acids joined by peptide bonds (Chothia 1976).
6. Residue volume (feature index 3.6, 4.6, and 5.6) - Bigelow described residue volume as the average hydrophobicity of proteins and the relation between it and protein structure (Bigelow 1967)
7. Steric Parameter (feature index 3.7, 4.7, and 5.7) – Charton described the steric parameter as values that define the tendency of amino acids to be found at the surface of globular proteins as a function of side-chain structure. Where globular protein tertiary structure has given the protein a rounded, globe like shape (Charton 1981).
8. Relative mutability (feature index 3.8, 4.8, and 5.8) - Dayhoff et al. described the relative mutability of each amino acid as the probability that amino acid will change over a small evolutionary period. The total number of changes are counted (on all branches of all protein trees considered), and the total number of occurrences of each amino acid is also considered (Dayhoff, Eck et al. 1972).

4.3.3.1 Normalized Moreau-Borto Autocorrelation

The first autocorrelation feature group is moreau-borto and is calculated using Eq 4-3.

$$AC(d) = \sum_{i=1}^{N-d} P_i P_{i+d} \quad \textbf{Eq 4-3}$$

where *d* is the lag of the autocorrelation d = 1, 2 ,3…30, P_i and P_{i+d} are the amino acid index property at position *i* and *i+d*, respectively and *N* is the length of the sequence. The lag relates to how many descriptor values are calculated for each autocorrelation subset feature, in the case of normalized moreau-borto autocorrelation method calculated in this study it was set to 30. Eq 4-4 is then normalised to obtain the normalised moreau-borto autocorrelation feature group, which is defined in Eq 4-4 (Li, Lin et al. 2006).

$$ATS(d) = \frac{AC(d)}{(N-d)} \quad \textbf{Eq 4-4}$$

Feature index 3 in Table 4-13 (if present) indicate the normalized moreau-borto autocorrelation feature group and index 3.1 to 3.8 (if present) are the sub features of the normalized moreau-borto autocorrelation feature group.

4.3.3.2 Moran Autocorrelation

The second autocorrelation feature group is moran autocorrelation defined in Eq 4-5.

$$I(d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}(P_i - \overline{\boldsymbol{P}})(P_{i+d} - \overline{\boldsymbol{P}})}{\frac{1}{N}\sum_{i=1}^{N}(P_i - \overline{\boldsymbol{P}})^2} \quad \textbf{Eq 4-5}$$

where d = 1, 2, 3, . . . , 30 is the lag of autocorrelation, P_i and P_{i+d} are the amino acid index property at position *i* and *i+d*, respectively and *N* is the length of the sequence, $\overline{\boldsymbol{P}}$ is the average of the considered amino acid index property ***P*** along the sequence.

$$\overline{\boldsymbol{P}} = \frac{\sum_{i=1}^{N} P_i}{N} \quad \textbf{Eq 4-6}$$

Where the sub feature group moran autocorrelation differs from the sub feature group moreau-borto autocorrelation is in the use of the amino acid index property deviations from the average values instead of the property values themselves as the basis for measuring correlations (Li, Lin et al. 2006).

Feature index 4 in Table 4-13 (if present) indicate the normalized moran autocorrelation feature group and index 4.1 to 4.8 (if present) are the sub features of the moran autocorrelation feature group.

4.3.3.3 Geary Autocorrelation

The third autocorrelation feature group is Geary Autocorrelation as defined in Eq 4-7.

$$C(d) = \frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}(P_i - P_{i+d})2}{\frac{1}{N-1}\sum_{i=1}^{N}\left(P_i - \overline{\boldsymbol{P}}\right)^2} \qquad \textbf{Eq 4-7}$$

where d = 1, 2, 3, . . . , 30 is the lag of autocorrelation, $\overline{\boldsymbol{P}}$ is the average of the considered property *P* along the sequence, P_i and P_{i+d} are the amino acid index property at position *i* and *i+d*, respectively and *N* is the length of the sequence. The Geary Autocorrelation algorithm compared to the others uses squared difference of amino acid index property values instead of vector product of amino acid index property values (Moreau-Borto Autocorrelation) or deviations (Moran Autocorrelation) as the basis for measuring correlations (Li, Lin et al. 2006).

Feature index 5 in Table 4-13 (if present) indicate the normalized geary autocorrelation feature group and index 5.1 to 5.8 (if present) are the sub features of the geary autocorrelation feature group.

4.3.4 Composition, Transition and Distribution

This feature group comprises of three types of features composition (C), transition (T) and distribution (D), each group contains seven structural and physicochemical properties, they are hydrophobicity, normalized Van der Waals volume, polarity, polarizibility, charge, secondary structures and solvent accessibility (Dubchak, Muchink et al. 1995; Dubchak, Muchnik et al. 1999; Cai, Han et al. 2003). These seven physicochemical properties are based on the clustering results of amino acid indices by Tomii and Kanehisa (Tomii and Kanehisa 1996). For each of the seven physicochemical properties, the amino acids are divided up into three classes such that those in a particular group are regarded to have approximately the same property values as shown in Table 4-4 (Li, Lin et al. 2006).

Table 4-4 Amino acid attributes for each physicochemical properties grouped into three classes

Physicochemical property (sub features)	Feature Index #	Classes		
		Class 1	**Class 2**	**Class 3**
Hydrophobicity	**6.1, 7.1, 8.1**	**Polar**	**Neutral**	**Hydrophobicity**
		R, K, E, D, Q, N	G, A, S,T,P,H,Y	C, L, V, I, M, F, W
Normalized Van der Waals volume	**6.2, 7.2, 8.2**	**Volume range 0–2.78**	**Volume range 2.95–94.0**	**Volume range 4.03–8.08**
		G, A, S, T, P, D	N, V, E, Q, I, L	M, H, K, F, R, Y, W
Polarity	**6.3, 7.4, 8.3**	**Polarity value 4.9–6.2**	**Polarity value 8.0–9.2**	**Polarity value 10.4–13.0**
		L, I, F, W, C, M, V, Y	P, A, T, G, S	H, Q, R, K, N, E, D
Polarizibility	**6.4, 7.4, 8.4**	**Polarizability value 0–1.08**	**Polarizability value 0.128–120.186**	**Polarizability value 0.219–0.409**
		G, A, S, D, T	C, P, N, V, E, Q, I, L	K, M, H, F, R, Y, W
Charge	**6.5, 7.5, 8.5**	**Positive**	**Neutral**	**Negative**
		K, R	A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V	D, E
Secondary structures	**6.6, 7.6, 8.6**	**Helix**	**Strand**	**Coil**
		E, A, L, M, Q, K, R, H	V, I, Y, C, W, F, T	G, N, P, S, D
Solvent accessibility	**6.7, 7.7, 8.7**	**Buried**	**Exposed**	**Intermediate**
		A, L, F, C, G, I, V, W	P, K, Q, E, N, D	M, P, S, T, H, Y

For each property, every amino acid is replaced by the index 1, 2 or 3 according to which of the three groups to which it belongs. Physicochemical properties such as charge (positive, negative and neutral) and secondary structure (helix, strand and coil) can only be divided into three classes while other physicochemical properties can be divided into many number of classes such as polarity and polarizibility as they are defined based numerical range (Li, Lin et al. 2006). For example, the protein amino acid sequence shown in Figure 4-1 will be encoded using the hydrophobicity physicochemical property shown in Table 4-4, the result of the encoding is shown in Figure 4-2. For the first amino acid residue ***S,*** its hydrophobicity physicochemical property class value equals to 2, subsequent amino acid residues ***T,A,G,K,V*** etc. class values are ***2,2,2,1,3*** etc.

>1HTB:A|PDBID|CHAIN|SEQUENCE

STAGKVIKCKAAVLWEVKKPFSIEDVEVAPPKAYEVRIKMVAVGICRTDDHVVSG
NLVTPLPVILGHEAAGIVESVGEGVTTVKPGDKVIPLFTPQCGKCRVCKNPESNYC
LKNDLGNPRGTLQDGTRRFTCRGKPIHHFLGTSTFSQYTVVDENAVAKIDAASPL
EKVCLIGCGFSTGYGSAVNVAKVTPGSTCAVFGLGGVGLSAVMGCKAAGAARIIA
VDINKDKFAKAKELGATECINPQDYKKPIQEVLKEMTDGGVDFSFEVIGRLDTMM
ASLLCCHEACGTSVIVGVPPASQNLSINPMLLLTGRTWKGAVYGGFKSKEGIPKLV
ADFMAKKFSLDALITHVLPFEKINEGFDLLHSGKSICTVLTF

Figure 4-1 Protein Sequence for protein 1HTB (Davis, Bosron et al. 1996)

22221331312233313112323113132221221313133232331211233221 3
322323332212223312321232231221133233221321313311212312311
132121223112211323121232233222232122331112321312222311333
323232222222313213222223233232232232332312222213323131111
321211322213312112112311331132122313231332131233223333212
322233323222211323123333221231223222312112321332133211323
123322332311311231332221233233 23.

Figure 4-2 Protein sequence 1HTB converted into hydrophobicity physicochemical property

The combined composition, transition and distribution is analysed as a whole feature group, feature index 6 in Table 4-13 (if present) is the result, and then each feature group is analysed separately as a sub feature group 6.1, 7.1 and 8.1.

4.3.4.1 Composition

Composition feature group defines the global percentage for each encoded class in a given protein amino acid sequence. Using the protein sequence example shown in Figure 4-1, the frequency of each class (class 1, class 2 and class 3) are 96, 147 and 131, respectively. The composition of each class is 96/374=25.67%, 147/374=39.30% and 131/374=35.03%, respectively, where 374 equals the given protein amino acid sequence length. The feature index 6.1 composition feature group formula is defined in Eq 4-8 (Li, Lin et al. 2006).

$$C_r = \frac{n_r}{N} \quad \text{r= 1, 2, 3} \qquad \textbf{Eq 4-8}$$

where N_r is the number of class *r* in the encoded protein sequence and *N* is the length of the protein sequence.

Feature index 6.1 Table 4-13 (if present) indicate the composition feature group and where feature index 6.1.1 to 6.17 (if present) indicate the sub features of the composition feature group results.

4.3.4.2 Transition

The transition feature group computes the transition between the different classes, i.e. from class 1 to class 2 or class 1 to class 3 or class 2 to class 3 and its vice versa class 2 to class 1, class 3 to class 1 and class 3 to class 2 in a given protein amino acid sequence. Using the protein sequence example shown in Figure 4-1 the number of times transition occurs between class 1 to class 2 is 71 times, class 1 to class 3 71 times and class 2 to class 3 is 108 times. These values are then converted into global percentage (in terms of sequence length), the percentages are class 1 to class 2 transition (71/373) = 19.03%, class 1 to class 3 transition (71/373) = 19.03% and class 2 to class 3 transition is (108/373) = 28.95%, where 373 equals the given protein amino acid sequence length (374-1) and the left hand side of the division in Eq 4-9 equals to the number of times a transition occurred between classes as defined in Table 4-4. The transition formula is defined in Eq 4-9.

$$T_{rs} = \frac{n_{rs} + n_{sr}}{N-1} \quad \text{rs= 12, 13, 23} \qquad \textbf{Eq 4-9}$$

where $n_{rs,}$ n_{sr} are the numbers of dipeptides encoded as *rs* and *sr* and *N* is the length of the protein sequence.

Feature index 7.1 Table 4-13 (if present) indicate the transition feature group and where feature index 7.1.1 to 7.17 (if present) indicate the sub features of the transition feature group results.

4.3.4.3 Distribution

This feature group calculates the distribution of each class in a given protein amino acid sequence at five specific positions. There are five sets of positions calculated along the sequence for each class (5*3 = 15 descriptor values) where 5 = number of positions and 3 =

number of classes. These positions are (1) the first position of class *r* residue, (2) 25% position of class *r* residue, (3) 50% position of class *r* residue, (4) 75% position of class *r* residue and (5) 100% position of class *N* residue.

Distribution of class 1, using the protein sequence example shown in Figure 4-1, positions of class 1 polar region amino acid residues (R, K, E, D, Q, N) for hydrophobicity sub feature are shown in Table 4-5. Six sub features can be calculated, for this example hydrophobicity sub feature will be used. The first class 1 (polar) amino acid residues to appear in the protein sequence is K, which is the 5th residue along the sequence. Reading Table 4-5 from left to right, the first cell contains number 5 which corresponds to the 5th residue in Figure 4-1 which is the K amino acid and second cell number in Table 4-5 is 8 which corresponds to the 8th residue in Figure 4-1 which is K amino acid residue and so forth.

Table 4-5 Positions of class 1 polar region amino acid residues (R, K, E, D, Q, N) based on Figure 4-1. Highlighted in bold are the 1st, 25%, 50%, 75% and 100% positions.

5	8	10	16	18	19	24	25	27	32
35	37	39	47	49	50	56	68	74	78
84	87	88	**96**	99	101	104	105	107	109
113	114	115	118	120	124	125	128	129	133
135	148	153	154	155	159	161	**167**	168	185
188	212	218	223	225	226	227	228	231	233
234	239	242	244	245	247	248	251	252	255
256	**259**	263	267	271	273	284	299	300	304
312	315	323	325	326	330	334	338	339	343
353	354	356	357	360	366	**369**			

The first position of a class 1 residue is the 5th residue (cell 1 in Table 4-5) , 25% position of a class 1 (C_r / 100) * 25 = 96th residue, 50% position of a class 1 (C_r / 100) * 50 = 167th residue, 75% position of a class 1 (C_r / 100) * 75 = 259th residue and 100% position of a class 1 (C_r / 100) * 100 = 369th residue, where C_r = 97 being the number of class 1 residues shown in Table 4-5.

The calculation of amino acid distributions for hydrophobicity sub feature class 1 is defined as (5/374)*100 = 1.34, (96/374)*100 = 25.67, (167/374)*100 = 44.65, (259/374)*100 = 69.25 and (369/374)*100 = 97.66, where left hand of the division is the first, 25%, 50%, 75% and 100%

position of class 1 residues and the right hand of the division sign is the length of the protein sequence. The values derived are the distribution values for class 1.

Distribution of class 2, using the protein sequence example shown in Figure 4-1 positions of class 2 neutral amino acid residues (G, A, S, T, P, H, Y) for hydrophobicity sub feature are shown in Table 4-6 is read from left to right; the first cell contains number 1, which corresponds to the 1st residue in Figure 4-1 which is the S amino acid, second cell number in Table 4-6 is 2 which corresponds to the 2nd residue in Figure 4-1 which is the T amino acid and so forth.

Table 4-6 Positions of class 2 neutral amino acid residues (G, A, S, T, P, H, Y) based on Figure 4-1. Highlighted in bold are the 1st, 25%, 50%, 75% and 100% positions.

1	2	3	4	11	12	20	22	29	30
31	33	34	42	44	48	51	54	55	59
60	62	66	67	69	70	71	75	77	79
81	82	85	86	91	**94**	95	98	106	108
110	117	119	121	122	126	127	131	134	136
138	139	142	143	144	145	147	149	150	156
158	162	163	164	165	173	175	177	178	179
180	181	**182**	183	187	190	191	192	193	194
196	199	201	202	204	206	207	210	213	214
215	216	217	221	230	232	236	237	238	243
246	249	258	260	261	265	270	274	277	**278**
283	285	287	288	289	293	295	296	297	298
302	305	310	311	313	316	317	319	320	321
324	327	329	333	337	341	344	347	348	351
358	363	364	365	367	370	**373**			

The first position of a class 2 residue is the 1st residue (Cell 1 in Table 4-6), the 25% position for class 2 (C_r / 100) * 25 = 94th residue, 50% position for class 2 (C_r / 100) * 50 = 182nd residue, 75% position for class 2 (C_r / 100) * 75 = 18th residue and 100% position for class 2 (C_r / 100) * 100 = 373rd residue, where C_r = 147, the number of class 2 residues shown in Table 4-6.
The calculation of amino acid distributions for class 2 is defined as, (1/374)*100 = 0.27, (94/374)*100 = 25.13, (182/374)*100 = 48.66, (278/374)*100 = 74.33 and (373/374)*100 = 97.73, the values derived are the distribution values for class 2.

Distribution of class 3 can be calculated the same way as class 1 and class 2, the sequence driven features derived for all three classes are shown in Table 4-7.

Table 4-7 Distribution of class 1, 2, 3 sequence driven features for the sub feature 8.1.1 hydrophobicity

Position	Class 1	Class 2	Class 3
1st	1.34	0.27	1.60
25%	25.67	25.13	23.80
50%	44.65	48.66	52.67
75%	69.25	74.33	77.45
100%	97.66	97.73	100

Feature index in 8.1 Table 4-13 (if present) indicate the distribution feature group and where feature index 8.1.1 to 8.7 (if present) indicate the sub features of the distribution feature group results.

4.4 Sequence Order

The sequence-order feature group relates to the distribution of amino acid residues in a protein's amino acid sequence. If a proteins order of amino acids residues changes, most likely, depending on how many changes and the differences between the amino acids have been made, it will change the shape of the protein and thus its function (Russell 2012). The Sequence Order feature group has two sub features (1) sequence-order-coupling numbers and (2) quasi-sequence-order descriptors (Li, Lin et al. 2006). Each sub feature is derived from both the Schneider-Wrede chemical distance matrix and the Grantham chemical distance matrix between each pair of the 20 amino acids (Li, Lin et al. 2006). In total, there are 160 descriptor values for the feature group and there sizes are shown in Table 4-8.

Table 4-8 Sequence order feature group and sub features descriptor size

Feature Index Number	Feature Name	Feature Size
9	Whole Sequence Order Group	160
9.1	Sequence-order-coupling number	60
9.1.1	Based on Schneider -Wrede distance	30
9.1.2	Based on normalized Granthan chemical distance	30
9.2	Quasi-sequence-order descriptors	100
9.2.1	Based on Schneider -Wrede distance	50
9.2.2	Based on normalized Granthan chemical distance	50

An illustrative example of sequence-order is shown in Figure 4-3 between two different protein sequences, sequence 1 is protein 1THB and sequence 2 is a random ordered sequence based on sequence 1. Both sequences have the same amino acid and dipeptide composition regardless of sequence-order as shown in Table 4-9 and Table 4-10 (each table only shows the

first 10-descriptor values). The sequence order is shown in Table 4-11 and the different order of amino acid results in two different sets descriptor values.

>sequence1 1HTB Human Alcohol Dehydrogenase protein

STAGKVIKCKAAVLWEVKKPFSIEDVEVAPPKAYEVRIKMVAVGICRTDDHVVSGNLVTPLPVILGHE
AAGIVESVGEGVTTVKPGDKVIPLFTPQCGKCRVCKNPESNYCLKNDLGNPRGTLQDGTRRFTCRGK
PIHHFLGTSTFSQYTVVDENAVAKIDAASPLEKVCLIGCGFSTGYGSAVNVAKVTPGSTCAVFGLGGV
GLSAVMGCKAAGAARIIAVDINKDKFAKAKELGATECINPQDYKKPIQEVLKEMTDGGVDFSFEVIGR
LDTMMASLLCCHEACGTSVIVGVPPASQNLSINPMLLLTGRTWKGAVYGGFKSKEGIPKLVADFMA
KKFSLDALITHVLPFEKINEGFDLLHSGKSICTVLTF

>sequence2 – Randomly generated sequence

PLEKVCLIGCGFSSIEKINEGFDLLHSGKEDVEVLTGRTWKGAVYGAPPKAYEVRIKMVAVGICRTDDH
CTGYVLPFSICTVVVSGNLVTPLPVILGHEAAGIVPLFTPQCGLGTSTFSQYKCRVHAVMGLTFCKNPE
SNYCLKNDLGNPRGTLQDGTRRFTCRGKPIHHFTVVDENAVAKIDAASSAVNVAKVTPGSTCAVFGL
GGVGLSGSTAGKVIKCKAAVLWEVAVDINKDKKPFKAAGAARIIKGDKVIFAKAKELGATECINPQDY
KKPIQESVGEGVTTVKPEVLKEMTDGGVDFSFEVIGRLDTMMASLLCCHEACGTSVIVGVPPASQNL
SINPMLLGFKSKEGIPKLVADFMAKKFSLDALIT

Figure 4-3 Two different protein sequences with the same amino acid composition and dipeptide

Table 4-9 the first 10 amino acid composition values for Sequence 1 and 2 (see Figure 4-3)

Sequence 1			Sequence 2	
Feature	**Descriptor Values**		**Feature**	**Descriptor Values**
Amino Acid Composition			**Amino Acid Composition**	
[F1.1.1.1]	8.28877		[F1.1.1.1]	8.28877
[F1.1.1.2]	4.278075		[F1.1.1.2]	4.278075
[F1.1.1.3]	4.545455		[F1.1.1.3]	4.545455
[F1.1.1.4]	5.080214		[F1.1.1.4]	5.080214
[F1.1.1.5]	4.278075		[F1.1.1.5]	4.278075
[F1.1.1.6]	10.160428		[F1.1.1.6]	10.160428
[F1.1.1.7]	1.871658		[F1.1.1.7]	1.871658
[F1.1.1.8]	5.882353		[F1.1.1.8]	5.882353
[F1.1.1.9]	8.55615		[F1.1.1.9]	8.55615
[F1.1.1.10]	7.754011		[F1.1.1.10]	7.754011

Table 4-10 the first 10 dipeptide composition values for Sequence 1 and 2 (see Figure 4-3)

Sequence 1		Sequence 2

Feature	Descriptor Values		Feature	Descriptor Values
Dipeptide Composition			Dipeptide Composition	
[F1.2.1.1]	1.340483		[F1.2.1.1]	1.340483
[F1.2.1.2]	0.268097		[F1.2.1.2]	0.268097
[F1.2.1.3]	0.268097		[F1.2.1.3]	0.268097
[F1.2.1.4]	0		[F1.2.1.4]	0
[F1.2.1.5]	0		[F1.2.1.5]	0
[F1.2.1.6]	0.80429		[F1.2.1.6]	0.80429
[F1.2.1.7]	0		[F1.2.1.7]	0
[F1.2.1.8]	0		[F1.2.1.8]	0
[F1.2.1.9]	1.340483		[F1.2.1.9]	1.340483
[F1.2.1.10]	0.268097		[F1.2.1.10]	0.268097

Table 4-11 The first 10 computed sequence-order values for Sequence 1 and 2 (see Figure 4 12)

Sequence 1			Sequence 2	
Feature	Descriptor Values		Feature	Descriptor Values
Sequence Order			Sequence Order	
[F2.1.1.1]	-0.082898		[F2.1.1.1]	-0.072193
[F2.1.1.2]	-0.050712		[F2.1.1.2]	-0.008403
[F2.1.1.3]	0.014657		[F2.1.1.3]	-0.009611
[F2.1.1.4]	-0.026517		[F2.1.1.4]	-0.005753
[F2.1.1.5]	-0.061593		[F2.1.1.5]	-0.042624
[F2.1.1.6]	-0.006673		[F2.1.1.6]	-0.036184
[F2.1.1.7]	-0.033106		[F2.1.1.7]	0.033221
[F2.1.1.8]	-0.032845		[F2.1.1.8]	-0.02618
[F2.1.1.9]	0.021632		[F2.1.1.9]	0.003778
[F2.1.1.10]	0.004643		[F2.1.1.10]	0.027183

Sequence-order-coupling

The dth rank sequence-order-coupling numbers is defined is Eq 4-10.

$$\tau_d = \sum_{i=1}^{N-d} (d_{i,\ i+d})^2 \qquad d = 1, 2, \ldots 30 \qquad \textbf{Eq 4-10}$$

where $d_{i,\ i+d}$ is the distance between the amino acids at positions *i* and *i+d* (Li, Lin et al. 2006).

Quasi-sequence-order

Type-1 quasi-sequence-order (first 20 descriptors) is defined in Eq 4-11 (Li, Lin et al. 2006).

$$Xr = \frac{f_r}{\sum_{r=1}^{20} f_r + w \sum_{d=1}^{30} \tau_d} \quad r = 1,2,3, \ldots 20 \qquad \textbf{Eq 4-11}$$

Type-2 quasi sequence order (last λ descriptors) is defined in Eq 4-12 (Li, Lin et al. 2006).

$$Xd = \frac{w\tau_{d-20}}{\sum_{r=1}^{20} + w\sum_{d=1}^{30}\tau d} \quad d\text{=}21, 22, 23, \ldots, \lambda \qquad \textbf{Eq 4-12}$$

where λ (lambda) reflects the effect of sequence order; f_r is the normalized occurrence for each amino acid type *i* and *w* is a weighting factor *w=0.1* (Li, Lin et al. 2006). In this study the λ = 30, which reflects the effect of sequence order at a 30 tier correlation level. Tier correlation level is further explained in Eq 4-18.

Feature index 9.1 in Table 4-13 (if present) indicate the whole sequence-order feature group and where feature index 9.1.1 to 9.2.2 (if present) indicate the sub features of the sequence-order feature group results.

4.5 Pseudo amino acid composition

Kuo-Chen Chou proposed the PseAAC feature group in order to include the sequence-order information of a protein along with its AAC descriptor values. According to (Chou 2005), a protein sequence can be represented by either discrete or sequential number mode. Amino acid composition is a discrete number mode; where there are twenty discrete descriptors representing the frequency of each amino acid type, however, states, this approach ignores the sequence order and length of a protein sequence. To include such information into a sequence-driven-feature a sequential number mode approach must be taken which uses the entire protein sequence to represent the given sample protein (Chou 2005). This approach gives us an unlimited number of possible combinations of sequence representations, which is near impossible to examine all. PseAAC contains more than the 20 discrete numbers, these extra numbers contain sequence order and length effects, the extra discrete numbers are called lambda (20 + λ) (Li, Lin et al. 2006; Li, Lin et al. 2006; Chen, Chen et al. 2008).

Studies such as (Xiao, Shao et al. 2006) extended the use of pseudo amino acid to protein structural class prediction and demonstrated that given the order of sequence along with amino acid composition achieved results, although this study was on a small dataset of 204 proteins, there was an increase from 68.1% (amino acid composition) to 89.7% (pseudo amino acid).

PseAAC has been used for the prediction of protein structural classes (Chen, Tian et al. 2006; Chen, Zhou et al. 2006; Xiao, Shao et al. 2006; Lin and Li 2007; Zhang and Ding 2007; Zhang, Ding et al. 2008). A recent study (Zhang, Ding et al. 2008) used datasets 204 and 1189,

reported an overall accuracy of 97% and 56.9%, respectively using leave-one-out cross validation test procedure, however the 204 dataset has a low sample size and contains very high homology between sequences, which is why the result is near 100%.

PseAAC descriptor is made up of a *20+ λ* descriptors in which the first 20 descriptors is the amino acid composition and the λ descriptors reflect the effect of the sequence order. The difference between quasi-sequence order descriptor and PseAAC, is the sequence-order-coupling number τ_d is now replaced by the sequence order correlation factors θ_λ as defined in Eq 4-13. The PseAAC equations are form the supplementary materials available at ProFEAT website http://jing.cz3.nus.edu.sg/cgi-bin/prof/prof.cgi and PseAAC webserver http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/.

$$\theta_\lambda = \frac{1}{N-\lambda} \sum_{i=1}^{N-\lambda} \Theta\,(R_i, R_{i+\lambda}) \qquad \textbf{Eq 4-13}$$

where θ_λ is the first tier correlation factor that reflects the sequence order between all of the *λ* most contiguous residues along the protein sequence *(λ=1…30)* and *N* is the number of amino acid residues. $\Theta(R_i, R_{i+\lambda})$ is the correlation and is given by Eq 4-14 (Li, Lin et al. 2006).

$$\Theta(R_i, R_{i+\lambda}) = \frac{1}{3}\{[H_1(R_j) - H_1(R_i)]^2 + [H_2(R_j) - H_2(R_i)]^2 + [M(R_j) - M(R_i)]^2\} \qquad \textbf{Eq 4-14}$$

where $H_1(R_i)$, $H_2(R_i)$ and $M(R_i)$ represent the hydrophobicity, hydrophilicity, and side-chain mass amino acid indices, of amino acid R_i, respectively (Li, Lin et al. 2006).

A protein sequence *P* has N amino acid residues as shown in Eq 4-15

$$p = R_1 R_2 R_3 R_4 R_5 \ldots R_N \qquad \textbf{Eq 4-15}$$

where R_1represents the residue at position 1 of the protein sequence *p* and so forth (Li, Lin et al. 2006).

The sequence order correlation in PseAAC introduces more correlation factors of physicochemical effects compared to the sequence order coupling number sub feature shown in Eq 4-11. For each amino acid type, the composition is calculated, which is the first part of the PseAAC vector space and is defined in Eq 4-16 and the second part of PseAAC is defined in Eq 4-17, which forms the sequence order part of the vector space.

$$Xr = \frac{f_r}{\sum_{r=1}^{20} f_r + w\sum_{d=1}^{30} \theta_j} \qquad r = 1,2,3, \ldots 20 \qquad \textbf{Eq 4-16}$$

$$Xd = \frac{w\theta_{d-20}}{\sum_{r=1}^{20} f_r + w\sum_{d=1}^{30} \theta_j} \qquad d = 21, 22, 23, \ldots, \lambda \qquad \textbf{Eq 4-17}$$

where λ (lambda) reflects the effect of sequence order and is fixes number; f_r is the normalized occurrence for each amino acid type *i* and *w* is a weighting factor *w=0.1* (Li, Lin et al. 2006). The sequence order and length effects (λ) can be calculated through a set of sequence-order correlation factors defined in Eq 4-18. The correlation factor is depended upon the value of λ. The lambda value is not fixed and within this study it is set at λ=30 which makes the PseAAC feature group 50 descriptors.

$$\begin{cases} \theta_1 = \frac{1}{N-1}\sum_{i=1}^{N-1} \Theta(R_i, R_{i+1}) \\ \theta_2 = \frac{1}{N-2}\sum_{i=1}^{N-2} \Theta(R_i, R_{i+2}) \\ \theta_3 = \frac{1}{N-3}\sum_{i=1}^{N-3} \Theta(R_i, R_{i+3}) \qquad (\lambda < N) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \theta_\lambda = \frac{1}{N-\lambda}\sum_{i=1}^{N-1} \Theta(R_i, R_{i+\lambda}) \end{cases} \qquad \textbf{Eq 4-18}$$

Where $\theta_1$is called the 1st tier correlation factor that reflects the sequence order correlation between all the most contiguous residues along a protein sequence (Shen and Chou 2008; Chou 2009) as illustrated in Figure 4-4. The 2nd tier θ_2 correlation factor reflects the sequence order correlation between all the most 2nd most contiguous residues along a protein sequence (Shen and Chou 2008; Chou 2009). The 3rd tier θ_3 correlation factor reflects the sequence order correlation between all the most 3rd most contiguous residues along a protein sequence (Shen and Chou 2008; Chou 2009). This continues up to λ=30 correlation factor reflects the sequence order correlation between all the most 30th most contiguous residues along a protein sequence and $J_{i,j}$ is a function of amino acids $R_i$andR_j. Figures 4-3 to 4-5 shows a graphical representation of the tiered correlation between residues in a protein sequence for the sequence order aspect of the pseudo amino acid composition feature group. These figures are adapted from the PseAAC webserver www.csbio.sjtu.edu.cn/bioinf/PseAAC Shen et al. 2008 (Shen and Chou 2008).

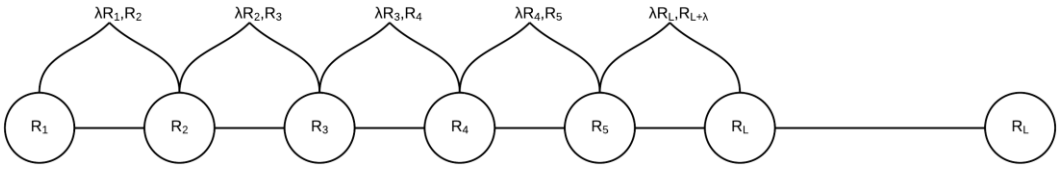


Figure 4-4 PseAAC 1st-tier correlation where $\lambda = 1$

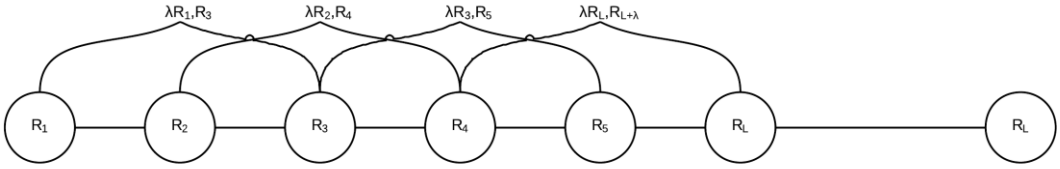


Figure 4-5 PseAAC 2nd-tier correlation where $\lambda = 2$

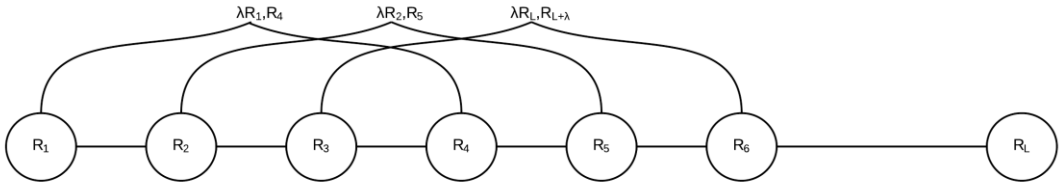


Figure 4-6 PseAAC 3rd-tier correlation where $\lambda = 3$

Feature index 10 in Table 4-13 (if present) indicates the PseAAC feature group and its sub features index 10.1 and 10.2 are the AAC and sequence-order (lambda) result.

4.6 Results and discussion

Table 4-13 contains the top 10 ranked results obtained through the analyses of sequence driven features for each dataset and test procedure. The results presented are ranked from the highest to lowest based on the overall percentage of predictive accuracy across all the four structural classes. The main analysis consisted of testing the full set of sequence-driven features over the four datasets. Each dataset was evaluated using three test procedures and classified using the MKNN algorithm. This approach produced 780 overall sets of results across

the 12 sets of dataset and test procedure combinations. The 780 overall ranked features are narrowed down to 120 (i.e., the top 10 results from each of the 12 combinations of dataset and test procedure). Results are focused on the top 10 results from each dataset and test procedure analysis, as it is more of an interest to discuss which sequence-driven features did well. The following sub sections present and discusses the results obtained from each sequence-driven feature group. The results shown in the subsequent tables are based on the highest achieved accuracies, which in all cases were obtained using the MKNN classifier. Each results table contains the results for each dataset and test procedure. Please refer to Appendix I for feature names. Where a specific set is referred to in the text, it relates to a specific dataset and test procedure analysis as shown in Table 4-12.

Table 4-12 Combination of datasets and test procedures

Set	Testing dataset	Test procedure
1	25PDB	10-fold
2	25PDB	Leave-one-out
3	25PDB	Independent-sets
4	1189	10-fold
5	1189	Leave-one-out
6	1189	Independent-sets
7	Astral25	10-fold
8	Astral25	Leave-one-out
9	Astral25	Independent-sets
10	Astral40	10-fold
11	Astral40	Leave-one-out
12	Astral40	Independent-sets

Table 4-13 Top 10 ranked features – feature index numbers are listed in Appendix I

Rank	Feature Index	MKNN neighbours	25PDB	Feature Index	MKNN neighbours	1189	Feature Index	MKNN neighbours	Astral25	Feature Index	MKNN neighbours	Astral40
10-fold												
1	1	[5,8]	41.69%	2	[1,2,5,9,10,11]	41.42%	6.1	[4,7,9,11]	41.40%	6.1	[5,6,8,9,10,11]	40.33%
2	6.1	[10,11]	41.55%	1	[3,5,10,11]	41.00%	1	[5,7,10,11]	39.45%	10	[3,8,10,11]	39.78%
3	6.1.6	[7,11]	39.03%	6.1	[5,10]	40.38%	10	[9,11]	38.13%	1	[8,11]	39.73%
4	10.1	[6,10,11]	37.77%	6.1.6	[1,9,10,11]	39.76%	10.1	[3,6,8,11]	37.76%	10.1	[5,10,11]	39.32%
5	610	[8,10]	37.65%	10.1	[3,5,8,9,10,11]	38.06%	2	[9,11]	37.15%	2	[8,9,10,11]	37.39%
6	11	[8,10]	36.33%	10	[3,11]	36.79%	6.1.6	[10,11]	36.75%	11	[9,10,11]	37.06%
7	2	[4,11]	35.87%	11	[3,4,8,9,10,11]	36.69%	11	[4,9,11]	36.25%	6.1.6	[4,5,8,9,10,11]	36.06%
8	3	[1,3,6,10,11]	35.33%	4	[9,11]	36.51%	6	[5,8,9,10,11]	34.71%	6	11	34.93%
9	6.1.1	[1,2,5,9,10,11]	35.14%	6	9	36.04%	3	11	34.60%	6.1.7	11	34.78%
10	4.1	11	34.36%	3	[3,5,6,7,8,9,10,11]	35.39%	6.1.7	10	34.60%	3	11	34.26%
Leave-one-out												
1	1	[5,8]	41.91%	2	[1,3,4,9,10,11]	41.84%	6.1	[8,11]	42.08%	1	[3,6,10,11]	42.63%
2	6.1	[4,7,8,9,11]	41.37%	1	[6,11]	41.75%	1	[9,11]	41.56%	6.1	[9,11]	41.57%
3	6.1.6	[7,11]	39.45%	6.1	[3,4,9,11]	40.46%	10.1	[6,8]	39.47%	2	[7,11]	40.74%
4	10.1	11	37.83%	10.1	[7,9]	39.54%	10	[9,11]	38.79%	10.1	[5,11]	39.63%
5	10	[3,11]	36.81%	6.1.6	[6,9]	38.89%	2	[8,9,10,11]	38.67%	11	[10,11]	39.58%
6	11	[8,9]	36.81%	10	[5,7,9,10,11]	38.62%	11	[9,11]	37.47%	10	[5,11]	38.96%
7	3	[3,8,10,11]	35.61%	4	9	37.51%	6.1.6	[4,9,10,11]	36.58%	6	[7,8,9,10,11]	37.23%
8	9.1.2	[4,8,9,11]	35.31%	3	[4,9,10,11]	36.68%	6	[9,11]	35.55%	3	[5,9,10,11]	37.14%
9	2	[8,11]	35.19%	11	[1,9,11]	36.31%	6.1.7	[8,9,10,11]	35.12%	6.1.6	[10,11]	36.95%
10	6.1.1	10	35.07%	6	[4,8,11]	35.02%	3	11	34.89%	4.1	[10,11]	35.76%
Independent-sets												
1	2	1	60.79%	2	1	63.87%	2	[1,2,3,5,10,11]	44.30%	1	[7,11]	41.38%
2	3	1	57.01%	5	1	61.29%	1	[1,2,3,8,10]	43.37%	2	[1,6,10,11]	40.27%
3	1	1	55.16%	4	1	60.18%	6.1	[1,4,8,10]	42.00%	6.1	[5,7,9,10]	40.16%
4	4	1	55.04%	3	1	59.72%	11	[1,2,4,5,7,11]	40.30%	11	[1,8,10,11]	38.50%
5	5	1	55.04%	1	[1,2,4,5]	57.88%	3	[1,2,3,10]	39.40%	6	[3,5,9,10,11]	37.01%
6	4.1	1	54.50%	11	1	55.85%	5.1	[1,2,4,8]	39.32%	6.1.6	[1,5,9,10,11]	37.01%
7	3.1	1	54.44%	4.1	1	55.12%	4	1	38.82%	10.1	[7,11]	36.40%
8	5.1	1	54.44%	5.1	1	54.10%	4.1	[1,2,7,8,11]	38.82%	3	[3,6,10,11]	36.09%
9	11	[1,2,3,10]	51.98%	3.1	1	53.83%	3.1	[1,2,3,4,11]	38.62%	5	[1,2,3,6,7,11]	35.35%
10	3.7	1	51.20%	9	1	53.36%	6	[1,2,5,10]	38.45%	10	[1,4,5,8,9,11]	35.32%

4.6.1 Results for amino acid composition feature group

The amino acid composition results are presented in Table 4-13 as feature index 1. Amino acid composition ranked the highest 4 out of 12 times using the 25PDB dataset with 10-fold test and leave-one-out test procedures, Astral40 dataset with leave-one-out and independent-sets test procedures. Amino acid composition also ranked in the top 10 with each dataset and test procedure analysis. The highest results across all datasets were achieved using the independent-set test procedure. Overall, the results show that the amino acid composition feature is one of the strongest predictor towards representing protein structural classes.

4.6.2 Results for dipeptide composition feature group

Dipeptide composition results are presented in Table 4-13 as feature index 2. Dipeptide composition ranked first 5 out 12 times using the 1189 dataset with 10-fold, leave-one-out and independent sets test procedures and 25PDB, Asral25 datasets the with independent-sets test procedures. The dipeptide composition is the best performing sequence-driven feature group, as it ranks in the top 10 with each dataset and test procedure analyses. The highest results across the 25PDB, 1189 and Astral25 datasets were achieved using the independent-set test procedure except the Astral40 where a marginally higher result was obtained using leave-one-out test procedure.

4.6.3 Results for autocorrelation feature groups

Autocorrelation results are presented in Table 4-13 with feature index 3 to 5.8. A separate table with only autocorrelation results is presented in Table 4-15 – this highlight the amino acid indices that frequently appear in the top 10 results. Feature index 3 represents the whole normalized moreau-borto autocorrelation feature group and its sub features, feature index 3.1 to 3.8. Feature index 4 represents the whole Moran autocorrelation feature group and its sub features, feature index 4.1 to 4.8. Feature index 5 represents the whole geary autocorrelation feature group and its sub features, feature index 5.1 to 5.8. The different autocorrelation feature groups differ in the way it derives the descriptor values a described in section 4.3.3. The autocorrelation sub features are amino acid indices representing various physicochemical and biochemical properties of amino acids.

Table 4-14 shows the selection of autocorrelation features across each dataset and test procedure analysis. The sub features that frequently appear in the top 10 ranked features are feature index numbers 3.1, 4.1 and 5.1, which are all the hydrophobicity scale (amino acid index CIDH920105) from each of the different autocorrelation feature group (feature index 3.1

is hydrophobicity scale from Normalized Moreau-Borto Autocorrelation, feature index 4.1 is calculated from Moran Autocorrelation and feature index 5.1 is calculated form Geary autocorrelation). The hydrophobicity scale was developed in the context of prediction of protein structural classes (Cid, Bunster et al. 1992) which explains why this sub features appears prominently within the results.

Table 4-14 the selection of autocorrelation features across all datasets and test procedures that appear in the top 10. Numbers in bold are autocorrelation sequence-driven feature groups. Numbers underlined and italicised are the hydrophobicity amino acid index sequence-driven feature (number denotes feature index number).

Test procedure	Datasets			
	25PDB	1189	Astral25	Astral40
10-fold	**3**	**4**	**3**	3
	4.1	**3**		
Leave-one-out	**3**	**4**	**3**	**3**
		3		*4.1*
Independent-sets	**3**	**5**	**3**	**3**
	4	**4**	*5.1*	**4**
	5	**3**	*4.1*	
	4.1	*4.1*	*3.1*	
	3.1			
	5.1	*5.1*		
	3.7	*3.1*		

Results from using the independent-sets test procedure, where the testing datasets are 25PDB and 1189 and training datasets are Astral25 and Astral40, respectively, 6 out of the top 10 ranked sequence-driven features are each of the whole autocorrelation sequence-driven feature groups followed by each autocorrelation's hydrophobicity sub feature. Results show that the hydrophobicity sub feature is always the next ranked feature behind its respective autocorrelation feature group. This shows that hydrophobicity amino acid index provides the majority of the predictive power for each of the autocorrelation sequence-driven feature groups. Overall, the best performing autocorrelation feature group is the Normalized Moreau-Borto autocorrelation; it ranked in the highest 9 out of 12 times. Normalised Moreau-Borto autocorrelation which has previously been used for predicting protein secondary structural contents (Lin and Pan 2001), the formula used to calculate the Normalized Moreau-Borto autocorrelation sequence-driven feature compared to the other two Autocorrelation sequence-driven feature group is a lot more simplistic. Although the observation is not scientific, it is merely a pattern seen that good results are derived using simple formulas as seen with using amino acid and dipeptide composition feature groups.

Table 4-15 Autocorrelation ranked features - feature index numbers are listed in Appendix I

Rank	Feature Index	MKNN neighbours	25PDB	Feature Index	MKNN neighbours	1189	Feature Index	MKNN neighbours	Astral25	Feature Index	MKNN neighbours	Astral40
10-fold												
8	3	[1,3,6,10,11]	35.33%	4	[9,11]	36.51%						
9							3	11	34.60%			
10	4.1	11	34.36%	3	[3,5,6,7,8,9,10,11]	35.39%				3	11	34.26%
Leave-one-out												
7	3	[3,8,10,11]	35.61%	4	9	37.51%						
8				3	[4,9,10,11]	36.68%				3	[5,9,10,11]	37.14%
9												
10							3	11	34.89%	4.1	[10,11]	35.76%
Independent-sets												
2	3	1	57.01%	5	1	61.29%						
3				4	1	60.18%						
4	4	1	55.04%	3	1	59.72%						
5	5	1	55.04%				3	[1,2,3,10]	39.40%			
6	4.1	1	54.50%				5.1	[1,2,4,8]	39.32%			
7	3.1	1	54.44%	4.1	1	55.12%	4	1	38.82%			
8	5.1	1	54.44%	5.1	1	54.10%	4.1	[1,2,7,8,11]	38.82%	3	[3,6,10,11]	36.09%
9							3.1	[1,2,3,4,11]	38.62%	5	[1,2,3,6,7,11]	35.35%
10	3.7	1	51.20%	3.1	1	53.83%						

4.6.4 Results for composition feature group

Composition results are presented in Table 4-13 where feature index number 6.1 is the whole composition feature group and its sub features have feature index numbers 6.1.1 – 6.1.7. Out of the 12 sets of the top 10 ranked features for each combination of dataset and test procedure analyses, over 25% of the top 10 ranked features come from the composition feature group. In each set of top 10 ranked features, excluding independent-sets where testing datasets are 25PDB and 1189, composition feature group (feature index 6.1) is closely followed (in terms of rank order) by its sub feature secondary structure (feature index 6.1.6). This shows that the predictive power of the composition feature group comes from the secondary structure sub feature. The secondary structure sub feature contains the composition of the three secondary structures, helix, strand and coil. It is one of these secondary structures that make up a gross majority structure content of a protein thus assigned its protein structural class based on that information. By removing the secondary structure sub feature (feature index 6.1.6) from the composition feature group, predictive accuracy decreases between 5% – 15% below that of the secondary structure sub feature, which shows further evidence that the majority predictive power of the composition feature group comes from the sub feature secondary structures, which is made up of only 3 descriptor values.

Table 4-16 Rank order of sequence feature index 6.1 composition feature group and index 6.1.6 sub feature secondary structure

25PDB		1189		Astral25		Astral40	
Rank	**Feature Index**	**Rank**	**Feature Index**	**Rank**	**Feature Index**	**Rank**	**Feature Index**
10-fold							
2	6.1	3	6.1	1	6.1	1	6.1
3	6.1.6	4	6.1.6	6	6.1.6	7	6.1.6
Leave-one-out							
2	6.1	3	6.1	1	6.1	2	6.1
3	6.1.6	5	6.1.6	7	6.1.6	9	6.1.6
Independent-sets							
				3	6.1	3	6.1
						6	6.1.6

4.6.5 Results for transition and distribution feature group

The transition between the classes defined in Table 4-4 which are based on pre-defined groupings and ranges by Li et al. (Li, Lin et al. 2006) and the distribution of amino acid residues at fixed points as defined in section 4.3.4.3 were not suitable with the selection of datasets used within this study. Both feature groups had no results ranked in Table 4-13, (where feature index numbers 7.1-7.1.7 or 8.1-8.1.7, transition or distribution feature groups, respectively), the sequence representations does not capture enough information to predict protein structural classes well enough.

However, these two sequence-driven feature groups does provide marginal predictive power when all three sequence-driven feature groups are combined (composition, transition and distribution) to form feature index 6 which appears in the top 10 ranked features 8 out of 12 times. In such a combined feature group, the predictive power still comes from sub features (feature index number 6.1.6) –secondary structure from the composition feature group. A trend has emerged from the results presented so far in that composition based sequence driven features (amino acid and dipeptide composition) better represent the prediction of protein structural classes than other types of sequence driven features.

4.6.6 Results for pseudo amino acid composition

Pseudo amino acid composition results are presented in Table 4-13 as feature index 10 to 10.2. With each dataset and test procedure analysis the PseAAC prediction result, except using Astral40 with independent-sets test procedure, was lower than what was achieved using the amino acid composition feature group. This rejected our initial theories that PseAAC would do better than traditional amino acid composition, which was based on the literature that including the effects of sequence-order information alongside amino acid composition includes the missing information that is not incorporated from solely using amino acid composition (Chou 2009).

This result led the way to explore why PseAAC did not do as well as amino acid composition, which was achieved by viewing the PseAAC results in three different angles. (a) as a whole feature group - with 50 descriptors (feature index 10), (b) only the amino acid composition aspect - the first 20 descriptor values (feature index 10.1) and (c) only the sequence order effect with the last $\lambda = 30$ descriptor values (feature index 10.2).

The sequence order effect (lambda) is not actually calculated into the amino acid composition aspect of the PseAAC feature group but exists as a separate set of descriptors, in which this study names as feature index 10.2. So then, it becomes possible to analyses the sequence order effect separately from the rest of the PseAAC feature group. The range of results between the three views and compared to amino acid composition are shown in boxplots. The boxplots compares the amino acid composition (feature index 1), pseudo amino acid composition feature group (feature index 10), the amino acid composition aspect of PseAAC sub feature (Feature 10.1) and the sequence order effect (lambda) (feature 10.2) averaged results for each test procedure across all datasets. Figure 4-7 shows the results for the 10-fold test procedure, Figure 4-8 shows the results for the leave-one-out test procedure and Figure 4-9 shows the results for the independent-sets test procedure.

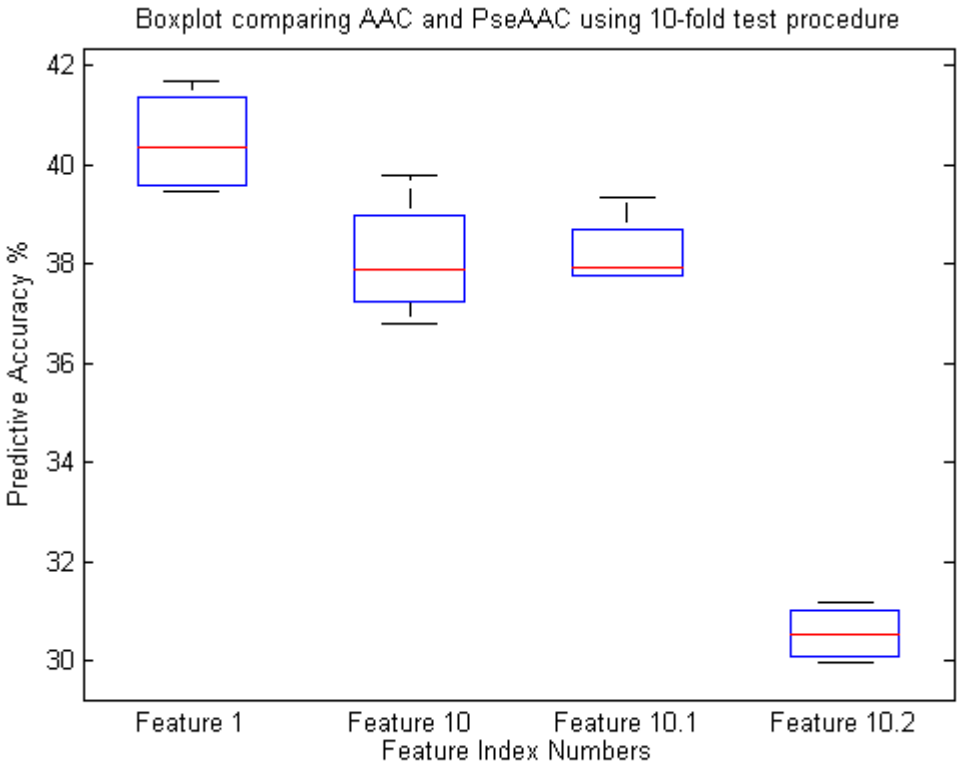


Figure 4-7 Boxplot for Pseudo amino acid composition for 10-fold test procedure (feature index numbers are listed in Appendix I) (Feature 1 amino acid composition, feature 62 PseAAC, feature 63 AAC part of PseAAC and feature 64 lambda part to PseAAC)

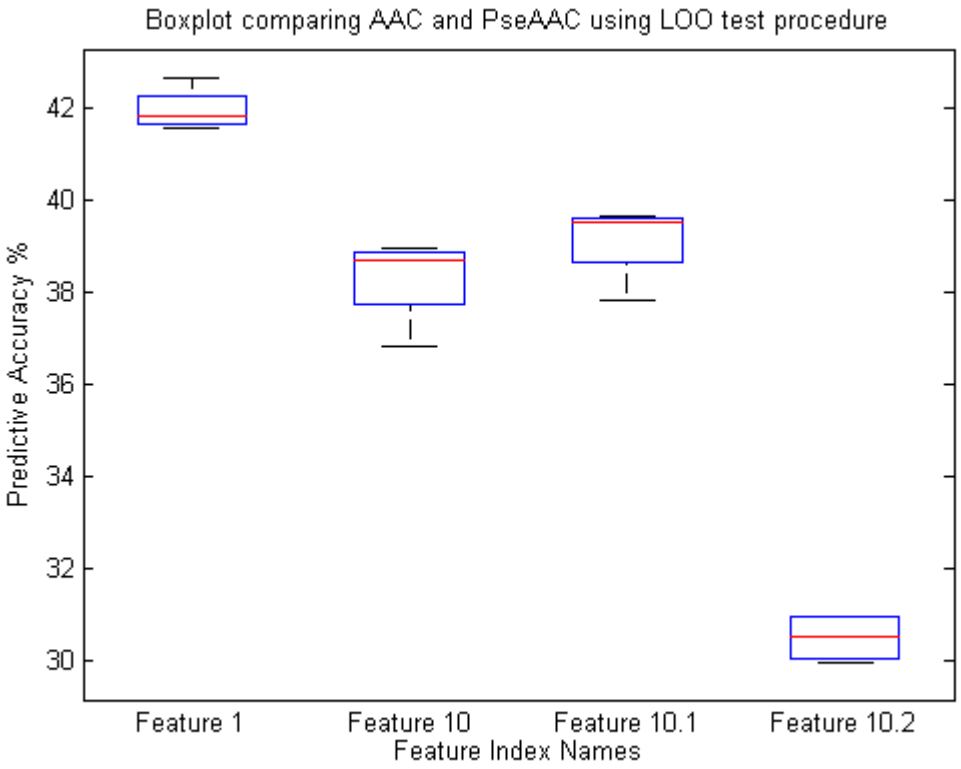


Figure 4-8 Boxplot for Pseudo amino acid composition for leave-one-out test procedure (feature index numbers are listed in Appendix I) (Feature 1 amino acid composition, feature 62 PseAAC, feature 63 AAC part of PseAAC and feature 64 lambda part to PseAAC)

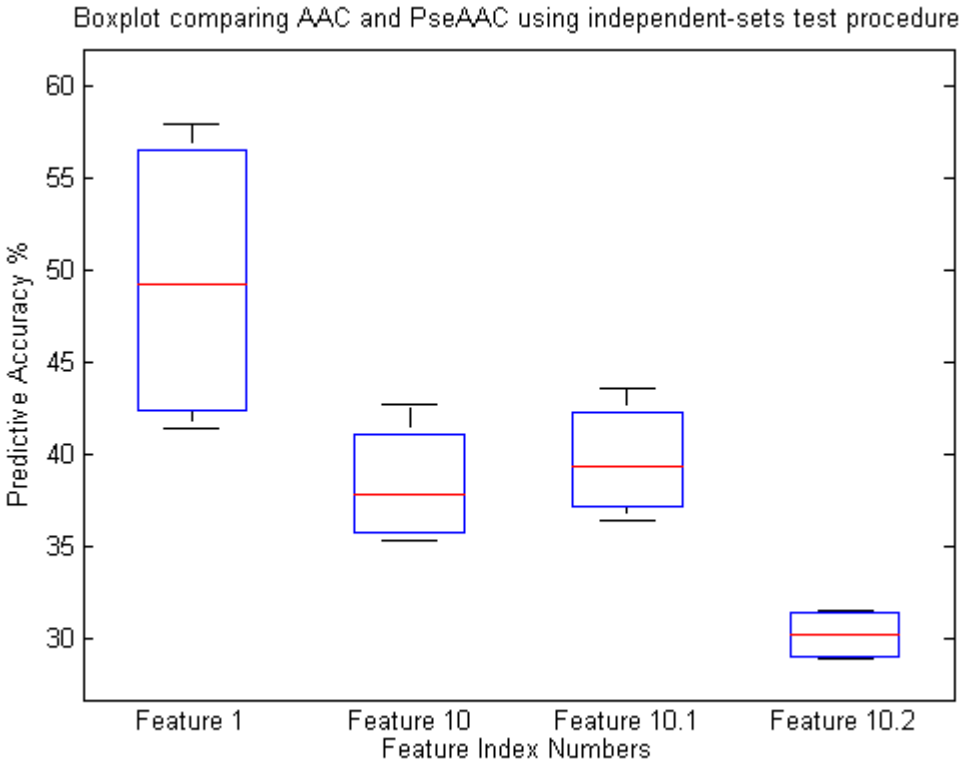


Figure 4-9 Boxplot for Pseudo amino acid composition for independent-sets test procedure (feature index numbers are listed in Appendix I) (Feature 1 amino acid composition, feature 62 PseAAC, feature 63 AAC part of PseAAC and feature 64 lambda part to PseAAC)

The view of results presented in the boxplots shows that the sequence-order effect (feature index 10.2), in the context of protein structural class prediction, does not add any significant increase in predictive accuracy to the overall pseudo amino acid composition (feature index 10). Each test procedure shows a clear trend how ineffective sub feature index 10.2 is by that it has the lowest range of results. When feature index 10.1 is combined with feature index 64 to make feature index 10, the range of predictive accuracies goes higher than using solely sequence order effect. However, it does not go higher than what has been obtained using traditional amino acid composition (feature index 1). Compared to feature index 63 it shows that the predictive power of PseAAC comes from the amino acid composition aspect of PseAAC. Table 4-17 compares the accuracies results of feature index 1, 10, 10.1 and 10.2, where a result is highlighted in bold mean the highest accuracy for the dataset and with test procedure. To conclude, PseAAC predictive power comes from its amino acid composition sub feature index 10.1 which essentially is traditional amino acid composition with a weighted parameter (w=0.1). Splitting the pseudo amino acid composition feature group and analysing

each sub feature separately allowed for a true representation of where PseAAC predictive power comes from. The results show that the amino acid composition part of PseAAC has a higher accuracy than PseAAC as a whole feature group, and even more so compared to the sequence order. The results are consistent across each dataset and test procedure analysis. This finding was not as expected, as recent studies (with low homology datasets) suggest that by incorporating sequence order effect (lambda) information with amino acid composition has a notable impact on improving the prediction quality (Chou 2001; Zhang, Ding et al. 2008). The weighted parameter has been shown to affect the result, albeit in a reduced fashion, leaves the question to be answered that a different weighted parameter may have an opposite effect on the result, this is an area that could be explored further.

4.6.7 Results of test procedures performance

One of the objectives of this study is to see how test procedures affect results across each dataset. Summary of the results across the entire feature set for each test procedure, (10-fold, leave-one-out and independent-sets) is shown in Figure 4-10, Figure 4-11 and Figure 4-12, respectively. These figures show the accuracies peaks of each feature group and sub features.

Table 4-17 Feature amino acid composition (feature index 1), PseAAC (feature index 10.2) , AAC of PseAAC (feature index 10.1) and PseAAC lambda (feature index 10.2) comparison - feature index numbers are listed in Appendix I

Feature Index	MKKN neighbours	25PDB	MKKN neighbours	1189	MKKN neighbours	Astral25	MKKN neighbours	Astral40
10-fold								
1	[5,8]	**41.69%**	[3,5,10,11]	**41.00%**	[5,7,10,11]	**39.45%**	[8,11]	39.73%
10	[8,10]	37.65%	[3,11]	36.79%	[9,11]	**38.13%**	[3,8,10,11]	**39.78%**
10.1	[6,10,11]	**37.77%**	[3,5,8,9,10,11]	**38.06%**	[3,6,8,11]	37.76%	[5,10,11]	39.32%
10.2	[3,4,6,8,9,11]	31.20%	[1,5,6,7]	30.24%	11	29.98%	[4,9,11]	30.84%
Leave-one-out								
1	[5,8]	**41.91%**	[6,11]	**41.75%**	[9,11]	**41.56%**	[3,6,10,11]	**42.63%**
10	[3,11]	36.81%	[5,7,9,10,11]	38.62%	[9,11]	**38.79%**	[5,11]	**38.96%**
10.1	11	**37.83%**	[7,9]	**39.54%**	[6,8]	39.47%	[5,11]	39.63%
10.2	[1,2,4,5,8,10]	30.94%	[7,8,9,11]	30.97%	11	30.09%	[6,9,10,11]	29.98%
Independent-sets								
1	1	**55.16%**	[1,2,4,5]	**57.88%**	[1,2,3,8,10]	**43.37%**	[7,11]	**41.38%**
10	[1,11]	39.51%	[3,9]	42.67%	[1,3,4,8,9,10,11]	36.14%	[1,4,5,8,9,11]	35.32%
10.1	[5,11]	**40.83%**	[5,11]	**43.59%**	[1,6,8,11]	**37.86%**	[7,11]	**36.40%**
10.2	[1,2,3,5,7,8]	31.54%	[7,11]	31.24%	7	28.88%	[1,4,9,10,11]	29.10%

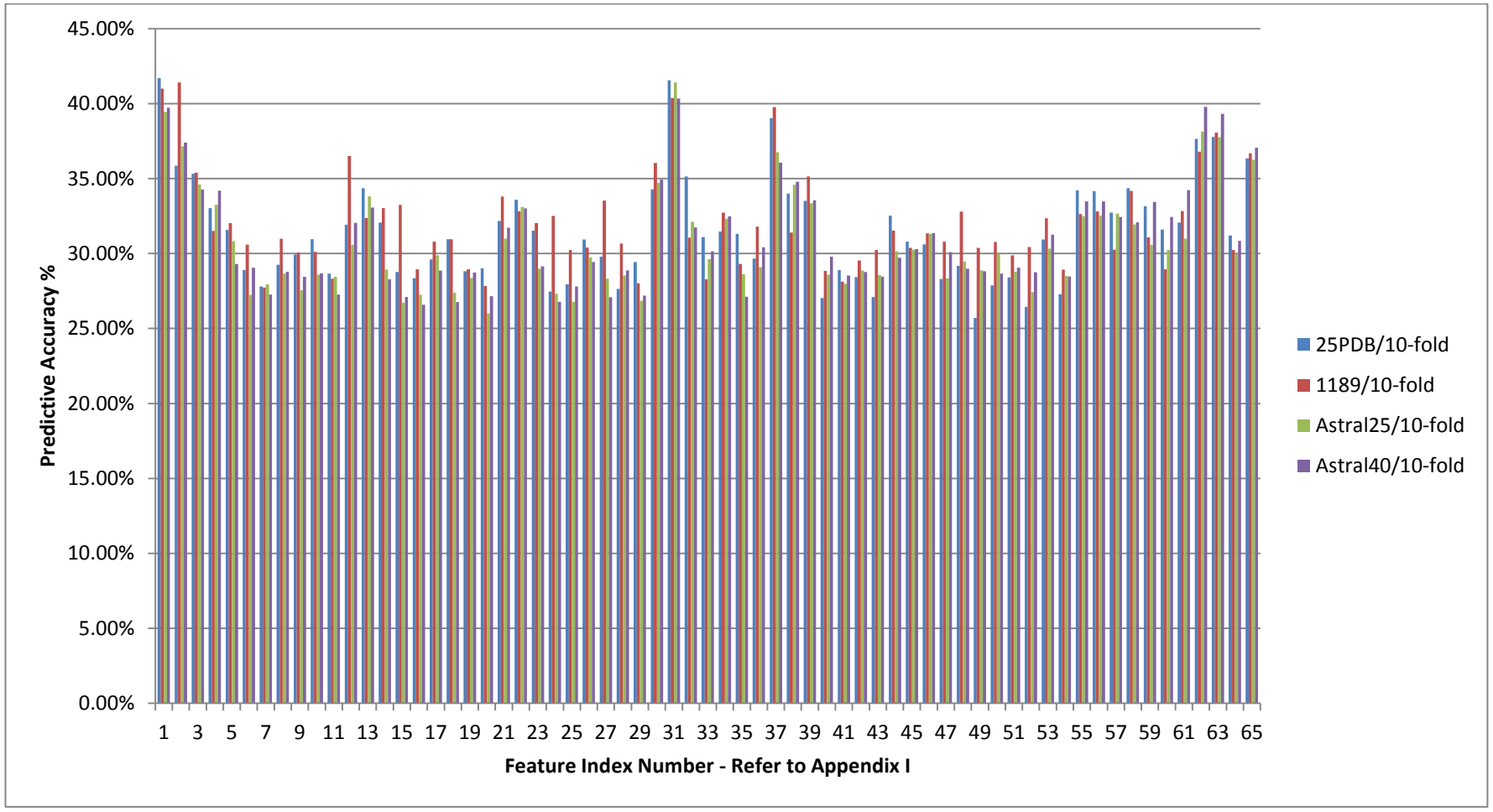


Figure 4-10 10-fold test procedure graphical view of each feature group and sub feature across each dataset - feature index numbers are listed in Appendix I

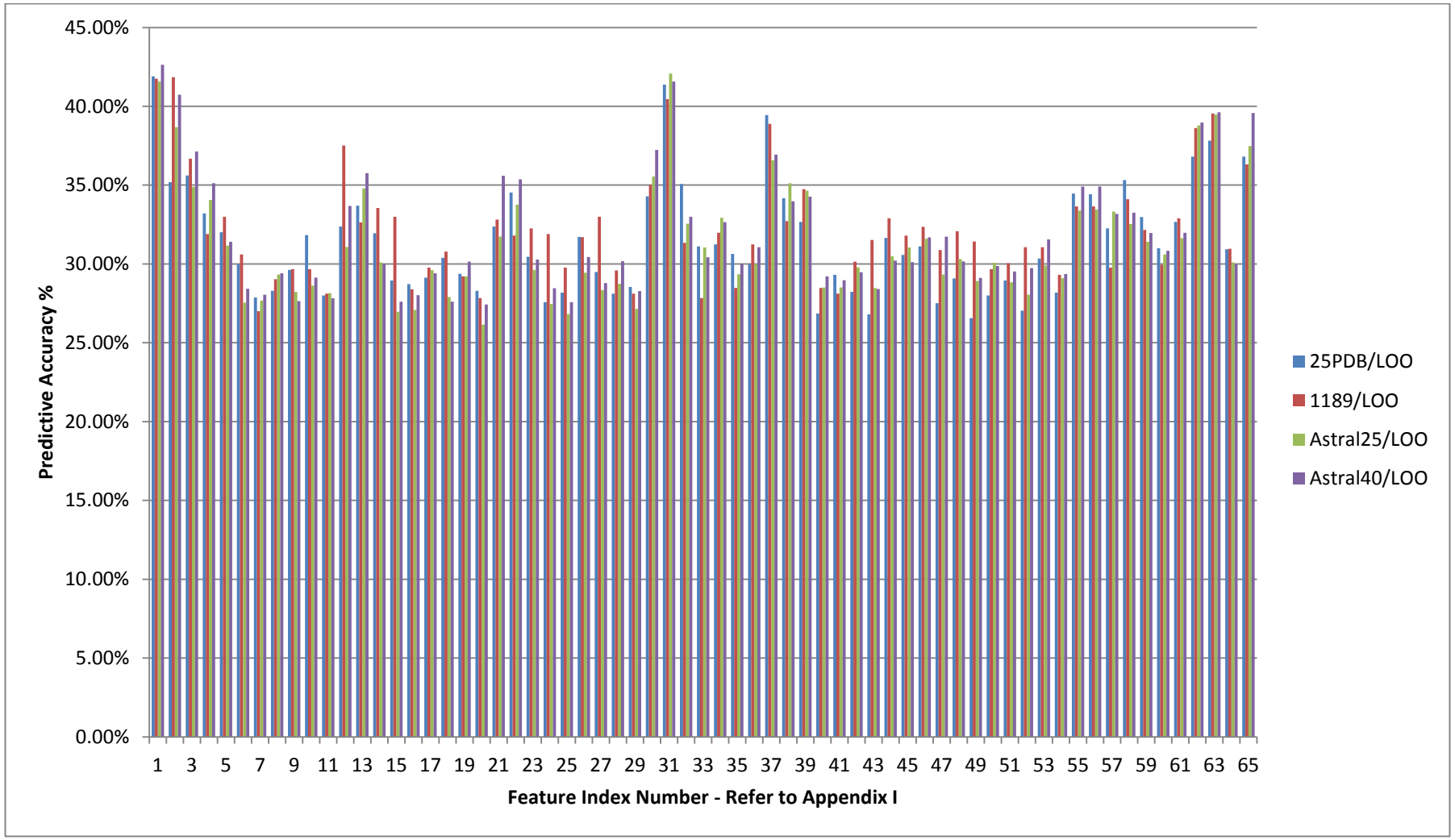


Figure 4-11 Leave-one-out test procedure graphical view of each feature group and sub feature across each dataset - feature index numbers are listed in Appendix I

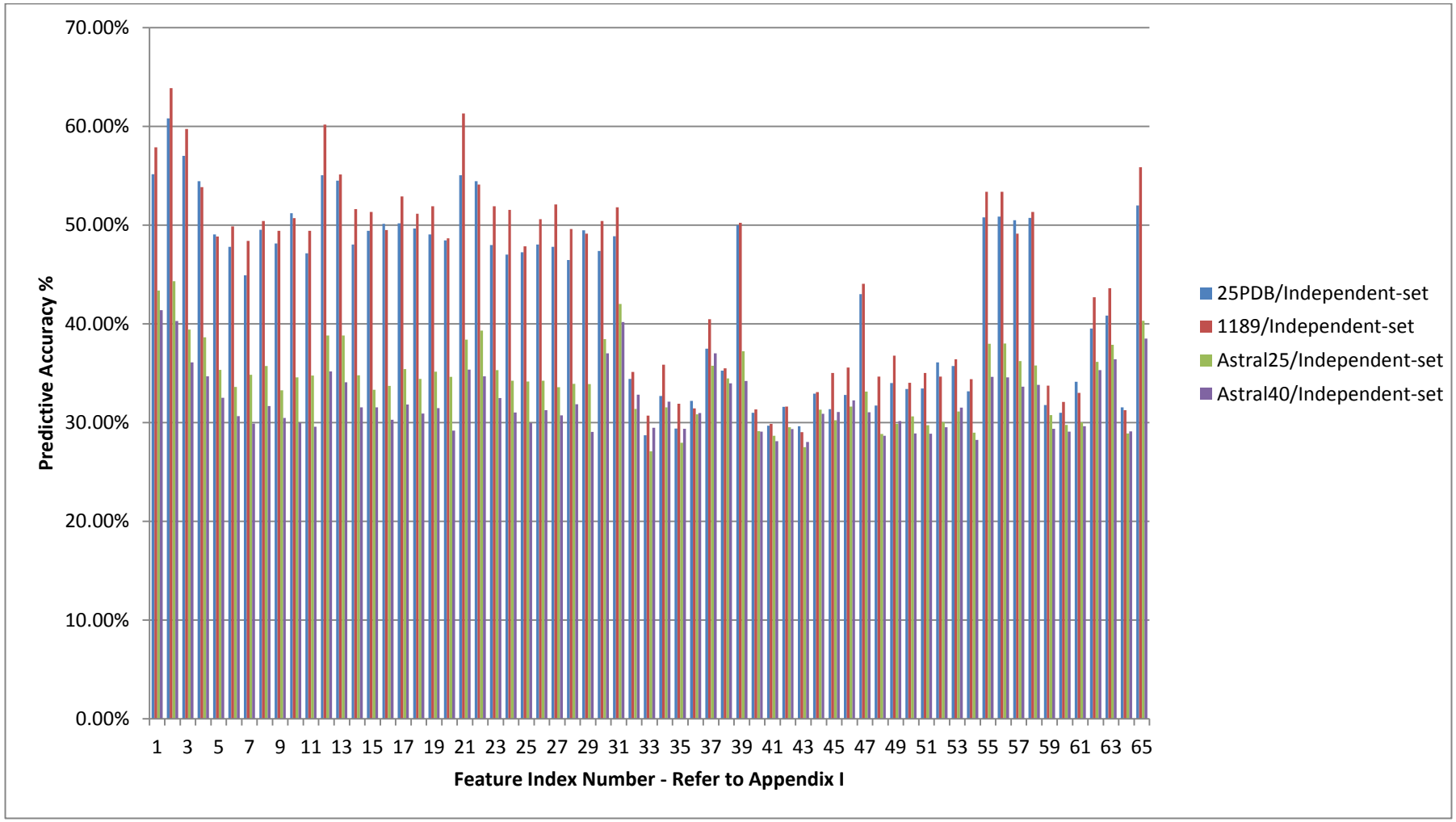


Figure 4-12 Independent-set test procedure graphical view of each feature group and sub feature across each dataset - feature index numbers are listed in Appendix I

Boxplot views of the range results across the entire feature set for each test procedure (10-fold, leave-one-out and independent-set) is shown in Figure 4-13, Figure 4-14 and Figure 4-15, respectively.

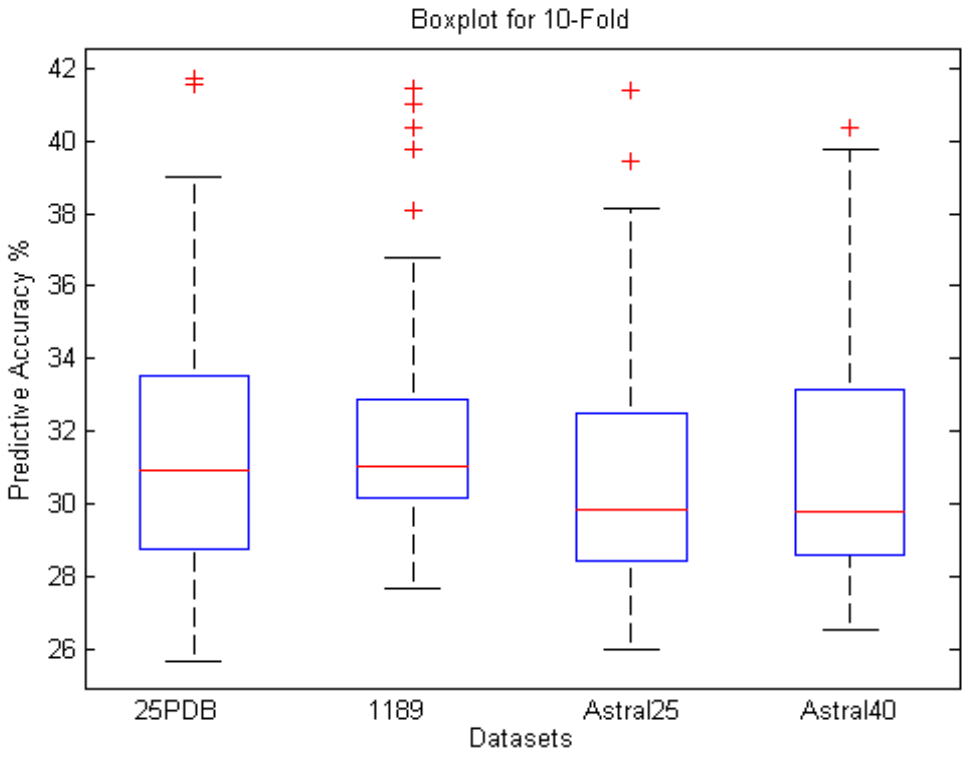


Figure 4-13 Boxplot for 10-fold test procedure

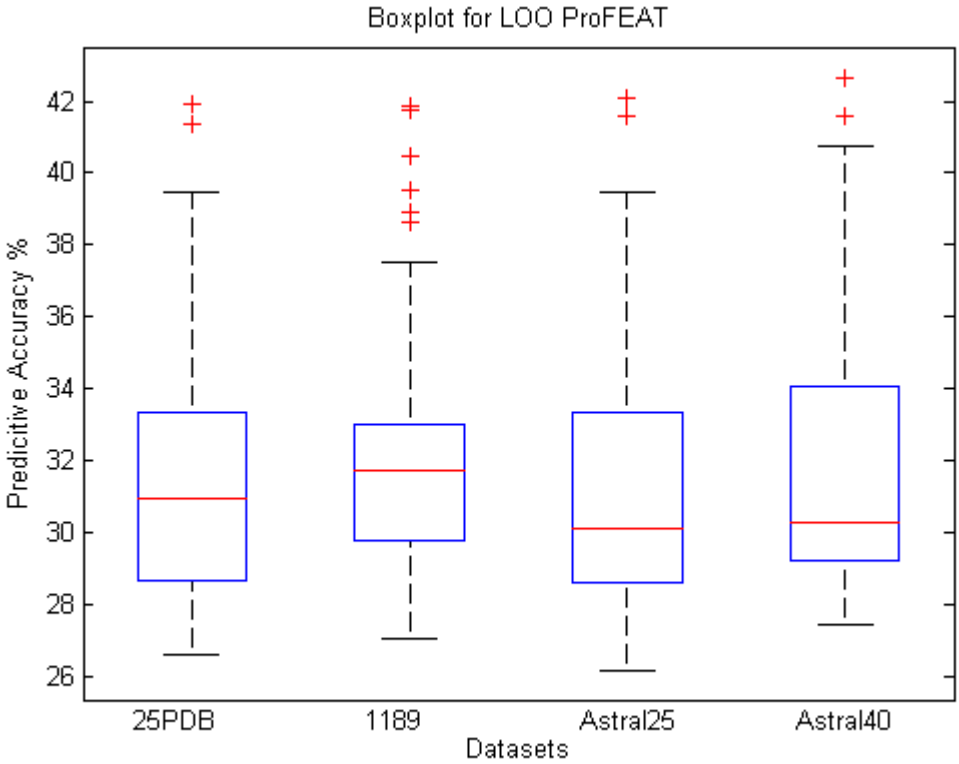


Figure 4-14 Boxplot for leave-one-out test procedure

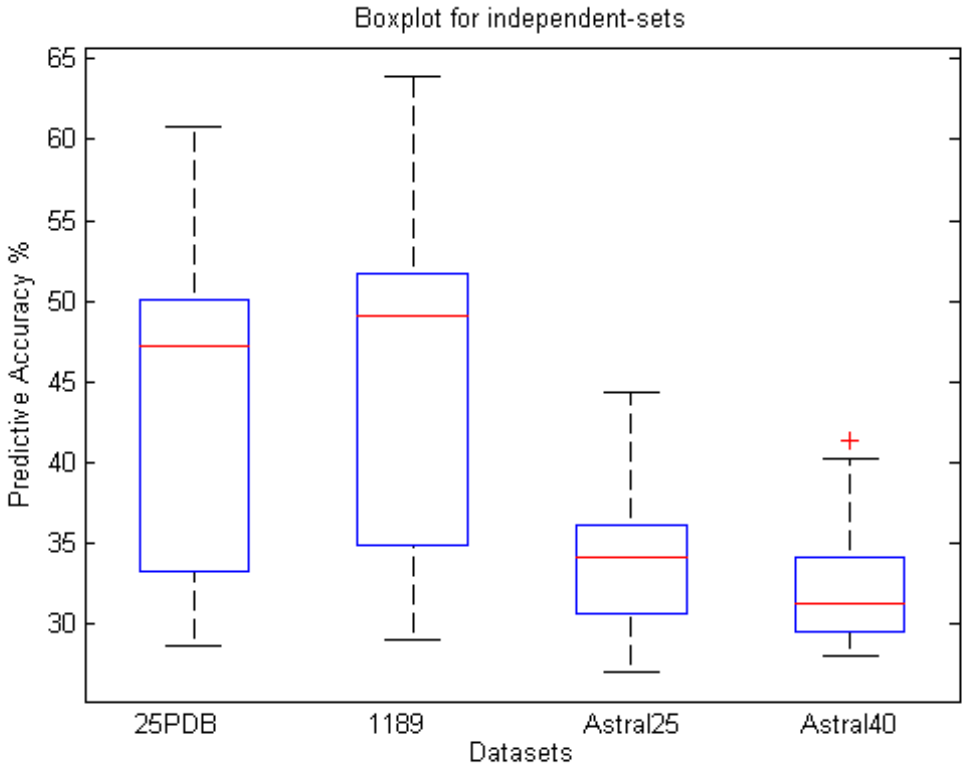


Figure 4-15 Boxplot for independent sets test procedure

The results show that test procedures affect the accuracies across the database using the same set of sequence driven features and the selection of features. 10-fold and leave-one-out test procedures output very similar set of results, with leave-one-out outputting a difference between -1.08% to 1.08% for the 25PDB dataset, -1.96% to 1.83% for the 1189 dataset, -0.42 to 2.12% for the Astral25 dataset and -2.26 to 3.88% for the Astral40 dataset. Compared to the independent sets test procedure where testing datasets are 25PDB and 1189, the difference increase from 10-fold is between -2.39% – 24.93% for the 25PDB dataset and is between -1.20% – 27.49% for the 1189 dataset. Compared to leave-one-out is between -2.40% to 25.60% and for the 25PDB dataset, is between -2.49% – 28.48% for the 1189 dataset. Results are more robust and achieve the highest accuracies by using independent-sets test procedure where the testing dataset are 25PDB and 1189 are trained on the larger Astral25 and Astral40 datasets, respectively. However, these results come at the expense at longer computational analysis times are needed to train the *mknn* classifier on the larger astral datasets.

The selection (or rank order) of features varies between 10-fold, leave-one-out and independent-sets test procedures. Much like how the accuracies range differs only slightly between 10-fold and leave-one-out (minimal differences), the selection of features also differs somewhat. Similarly, accuracies obtained through independent-sets are a lot higher than 10-fold and leave-one-out, the selection of features are more different. Amongst the variations in selected features across each test procedure the only 3 common features selected across each test procedure, that are ranked in the top 10 results are (1) amino acid composition – feature index 1, (2) dipeptide composition - feature index 2 and (3) Normalized Moreau-Borto Autocorrelation - feature index 3. This adds further weight into that composition based features are the most representative of protein structural classes and progress should be made using this type of sequence driven features. These differences and commonality show that each test procedure produces a different set of outcomes and that it is necessary to evaluate datasets using many different test procedures to derive a consensus set of results and not rely on a single test procedure.

4.6.8 Individual class performance

Examining individual structural class performance helps better identify which features are better at predicting specific protein structural classes. The results presented so far are based on the average accuracy across all four classes. From each set of results, the highest accuracy for each structural class is extracted. The selection of features presented Table 4-18, Table 4-19, Table 4-20 and Table 4-21 correlates with the features presented in section Table 4-13, with the exception of feature index 12 and 31. Feature index 1, 2, 3, 6.1 and 6.1.6 all composition based features and is the consensus set of features that represents individual classes. Full set of individual class results for each dataset and test procedure analysis are presented in Appendix VIII chapter 4 individual structural class of proteins results.

Table 4-18 Individual class majority selected feature 25PDB dataset /each test procedure

Structural Class	Dataset					
	25PDB					
	Test Procedures					
	10-fold		Leave-one-out		Independent-sets	
	Feature #	Accuracy	Feature #	Accuracy	Feature #	Accuracy
All-α	6.1.6	58.4%	6.1.6	59.5%	3	58.6%
All-β	6.1.6	53.03%	6.1.6	54.2%	2	62.59%
α/β	1	37.54%	1	37.21%	2	65.99%
α+β	3	40.83%	3	38.55%	2	61.68%

Table 4-19 Individual class majority selected feature 1189 dataset /each test procedure

Structural Class	Dataset					
	1189					
	Test Procedures					
	10-fold		Leave-one-out		Independent-sets	
	Feature #	Accuracy	Feature #	Accuracy	Feature #	Accuracy
All-α	3	48.04%	3	52.02%	4	60.54%
All-β	6.1	51.67%	1	54.79%	2	67.47%
α/β	2	63.64%	2	65.15%	2	70%
α+β	2	24.58%	4	26.25%	5	63.33%

Table 4-20 Individual class majority selected feature Astral25 dataset / each test procedure

Structural Class	Dataset					
	Astral25					
	Test Procedures					
	10-fold		Leave-one-out		Independent-sets	
	Feature #	Accuracy	Feature #	Accuracy	Feature #	Accuracy
All-α	31	44.89%	31	46.82%	3	56.84%
All-β	37	47.89%	31	49.37%	2	53.22%
α/β	3	37.12%	2	58.89%	2	47.38%
α+β	2	52.10%	3	37.56%	3	37.68%

Table 4-21 Individual class majority selected feature Astral40 dataset / each test procedure

Structural Class	Dataset					
	Astral40					
	Test Procedures					
	10-fold		Leave-one-out		Independent-sets	
	Feature #	Accuracy	Feature #	Accuracy	Feature #	Accuracy
All-α	3	42.28%	3	47.03%	3	54.82%
All-β	6.1.6	48.03%	6.1	48.67	1	54.52%
α/β	2	52.39%	2	62.86%	2	67.36%
α+β	3	36.81%	3	38.11%	6.1	67.36%

4.7 Conclusions

This chapter has presented a comprehensive analysis of the largest set of sequence-driven features for the prediction of protein structural classes. This detailed investigation explored the features and identified which are more suited at predicting PSC, we consider the results presented as a benched mark to compare future work. These sequence-driven-feature groups have been applied in many published studies, including protein structural classes; so far, individual group of features have been separately used in many protein structural class prediction studies, however, not all of them have ever been applied in a single protein structural class study. It was the aim of this chapter to examine how the use of different feature groups affects the performance of this and to develop a benchmark set of results.

Four datasets were used, two of which contain no more than 25% homology and two contained no more than 40% homology. The two datasets named Astral25 and Astral40 are the largest datasets ever constructed for the predication of protein structural classes. The other two datasets 25PDB and 1189 are widely used as benched mark datasets (Kurgan and Homaeian 2006).

The MKNN classifiers results are given in this chapter, which is the result of combined multiple analysed K's. Three different test procedures were used to evaluate MKNN's effectiveness. Test procedures are the standard way of evaluating a classifiers performance. The most common methods are N-fold cross-validation and leave-one-out; the chapter also tested the effects of using independent-sets, which is not widely used. It was identified through literature review that there has not been a study that investigated the effects of more than one test procedure as the more commonly used test procedures used were 10-fold or leave-one-out. It was found that each test procedure produced a different set of predictive accuracies and rank order of sequence-driven-features. Between 10-fold and leave-one-out results the rank order are different in some features but on a general overview the more important features groups such as amino acid composition and dipeptide remain consistently strong across all datasets with a very similar rank order across both test procedure. Independent-sets produced a different rank order of features compared to 10-fold and leave-one-out. This shows that 10-fold and leave-one-out are very similar, 10-fold is faster and leave-one-out is a more thorough but slower as it tests each protein sample individually, but gives slightly higher predictive accuracies than 10-fold test procedure. Independent-sets results are a lot different from 10-fold and leave-one-out; the main difference is where 25PDB

and 1189 are the testing datasets, the results increase up to 29% higher by training 25PDB and 1189 with Astral25 and Astral40, respectively.

The decision was taken to focus on the top 10 results for each sequence driven feature and sub-features to manage the large collection of analyses results. Focussing on the top 10 results also eliminates the weaker results. The highest-ranking sequence-driven-features are amino acid composition - feature index 1, dipeptide composition - feature index 2 and composition feature group - feature index 6.1. This gives strong evidence towards the best type of sequence driven features that are suited for the prediction of protein structural classes are composition based ones. Composition based sequence driven features are calculations of the composition of certain types of amino acids, or paring of amino acids or types of groups of amino acids.

It has been shown that results are highly dependent on selection of dataset and test procedures. No single combination of dataset or test procedure should be relied upon solely because results vary over different selection of dataset and test procedure. It is important to obtain a consensus set of results for a much better understanding how each combination of dataset and test procedure reacts to the same set of features. In addition, no single feature was capable of predicting each dataset and all its classes close to 100%, however it was narrowed down to composition based features that predicted protein structural classes better.

Further investigation into the sequence driven features groups revealed that the sequence driven feature groups autocorrelation and the PseAAC utilises amino acid indices to derive its descriptor values. Autocorrelation and PseAAC feature groups utilises eight and three amino acid indices, respectively. Amino acid indices represent many different physiochemical or biochemical properties of amino acids. It was identified that the amino acid indices utilised are a small selection of indices from the AAIndex1 database (Kawashima, Pokarowski et al. 2008), which contains over 500 amino acid indices of varying properties which have not been taken in to consideration. The results obtained from these feature group were amongst the top 10 ranked features across each dataset using the independent-set analysis – in particular feature index 3.1, 4.1, and 5.1, each of which each is derived from the hydrophobicity amino acid index (Cid, Bunster et al. 1992). However, since composition based feature groups resulted in the highest accuracies, this gave way to the idea of developing a sequence driven feature that is capable of utilising all of the indices available from the AAindex1 database acid index. The

composition based feature group that can be adapted to consider all the amino acid indices is the amino acid composition feature group. AAC assumes that each amino acid has a weight of one, however it may be possible to replace this artificial amino acid weight with a natural amino acid weight in the form of amino acid indices as there many different physiochemical or biochemical properties of amino acids.

Chapter 5, takes forward the finding that composition based sequence drive features are better suited to the prediction of protein structural classes and that amino acid indices to represent different physiochemical properties of amino acids provides a flexible approach of utilising different properties. A novel generalisation of amino acid composition feature group by utilising these amino acid indices as potential sequence driven features is presented in chapter 5.

Chapter 5 - Amino acid indices based sequence driven features

5.1 Introduction

The study presented in chapter 4 investigated the largest set of traditional sequence-driven-features and some of the sequence driven features have used a very small set of the amino acid indices. The sequence-driven-feature group PseAAC uses three different types of amino acid indices whereas the autocorrelation feature subsets utilises eight amino acid indices to derive its descriptor values. Amino acid indices are useful for better characterising amino acid properties (Huang, Kawashima et al. 2007; Georgiev 2009) and there is a database named the AAindex (Kawashima, Pokarowski et al. 2008) that contain many hundreds of them. Although the amino acid indices used within the autocorrelation and PseAAC sequence driven feature groups were not as strong as composition-based features, the idea of utilizing all available amino acid indices, not just a select few, requires further investigation. Chapter 5 looks at the usability of amino acid indices for the prediction of structural classes of protein, resulting in:-

1. Comprehensive analysis into amino acid indices for the prediction of structural classes of proteins
2. A hybrid method of clustering and PCA to remove redundant amino acid indices and generate novel indices
3. A novel generalised amino acid composition (GAAC) sequence driven method, which expands the current amino acid composition formula by utilising weights
4. GAAC webserver
5. Two novel feature extraction methods based on the mean and PCA

5.2 Amino acid indices

Proteins are defined by a combination of twenty naturally occurring amino acids along the protein sequence; these amino acids have been investigated through many experimental and theoretical studies since the early sixties. Amino acids properties can be represented by a fixed set of twenty descriptor values, which are known as an amino acid index representing a

certain physiochemical or biochemical property of a protein. An example of an amino acid index is shown in Table 5-1, it show's an amino acid index with ID ANDN92010 - alpha-CH chemical shifts (Andersen, Cao et al. 1992) in its raw and normalised format.

Table 5-1 Amino acid index ANDN920101 - alpha-CH chemical shifts from the Amino Acid Index Database

Selected Criteria [Hide]																			
name	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
ANDN920101-raw	4.35	4.38	4.75	4.76	4.65	4.37	4.29	3.97	4.63	3.95	4.17	4.36	4.52	4.66	4.44	4.50	4.35	4.70	4.60	3.95
ANDN920101-scaled	-0.27	-0.15	1.33	1.37	0.93	-0.19	-0.51	-1.79	0.85	-1.87	-0.99	-0.23	0.41	0.97	0.09	0.33	-0.27	1.13	0.73	-1.87

In 1988 Nakai et al collected 222 amino acid indices from published literature (Nakai, Kidera et al. 1988), in 1996 Tomii et al further extended the collection of amino acid indices to 402 entries (Tomii and Kanehisa 1996) and between 1996 to 2008 the collection has been updated by Kawashima et al (Kawashima, Pokarowski et al. 2008) to 544 amino acid indies. These amino acid indices are in the AAindex1 database http://www.genome.jp/aaindex at the GenomeNet in Japan (Kawashima, Pokarowski et al. 2008).

Different types of studies have utilised amino acid indices such as protein structural classes (Huang, Kawashima et al. 2007; Seker 2008; Nanuwa, Dziurla et al. 2009), protein subcellular location (Sarda, Chua et al. 2005), protein secondary structure (Kazemian, Moshiri et al. 2007; Kurgan, Stach et al. 2007), protein transmembrane sequences (Zhao and London 2006), protein chemical structure and biological function (Sneath 1966), protein surface prediction (Nishikawa and Ooi 1980; Nishikawa and Ooi 1986) and protein disordered regions (Han, Zhang et al. 2009).

Some of the above-mentioned studies used a limited number of amino acid indices, which is further evidence that there was scope for a comprehensive analysis into the role of amino acid indices for the prediction of protein structural classes. With an in-depth investigation, the identification of a candidate set of amino acid indices that are capable of predicting specific proteomic characteristic may leverage the prediction accuracy.

5.3 Amino acid indices database

The AAIndex1 database currently contains 544 amino acid indices; each consists of an accession number, a short description of the index, the reference information and the numerical values for the properties of 20 amino acids. All of the these indices are clustered into five main groups where the correlation coefficient of 0.9 is used as similarity threshold,

they are, alpha and turn propensities, beta propensity, composition, hydrophobicity, physicochemical properties and other properties, essentially these groups contain similar or related indices.

The amino acid indices dataset was downloaded for pre-processing, out of the 544 amino acid indices, thirteen of them: - AVBF000101, AVBF000102, AVBF000103, AVBF000104, AVBF000105, AVBF000106, AVBF000107, AVBF000108, AVBF000109, YANJ020101, GUYH850103, ROSM880104 and ROSM880105 were found to have missing index values. In addition, three amino acid indices RICJ880102, PRAM900102 and LEVM780102 have identical set of index values with RICJ880101, LEVM780101 and PRAM900101, respectively. Subsequently, these sixteen amino acid indices were removed from this study's amino acid indices dataset. At this point, the amino acid dataset consists of 528 amino acid indices.

The amino acid indices database AAIndex1 was last updated March 31st 2008, which contained indices up to 2006, this study was undertaken in mid-2009, in which three years had passed with no update. This led to literature searches for more amino acid indices which yielded an additional 83 from the following publications (Wilkins, Gasteiger et al. 1999; Atchley, Zhao et al. 2005; Zviling, Leonov et al. 2005; Fernandez, Caballero et al. 2007; Huang, Kawashima et al. 2007; Kurgan, Stach et al. 2007; Seker 2008; Asakawa, Sakiyama et al. 2010) bringing the total number of unique amino acid indices into the new dataset is 611, which best to our knowledge is the largest collection of amino acid indices. Full list of amino acid indices are in listed in appendix II.

5.3.1 Normalisation of amino acid indices

The final collection of amino acid indices originated from many different sources and each source has a different method of deriving indices values, which is dependent on the authors' calculation method. The different calculation methods can make the final set of index values have different ranges i.e. AAI 1 and AAI 2 are derived from two different sources, their index values are between 3.95-4.76 and 0-2.65, respectively. The amino acid indices dataset contains many more different ranges and for consistency each amino acid index is normalised using z-score function (Marx 2006) as shown in Eq 5-1where E, $\bar{x}$ and σ correspond to index value, mean value and standard deviation for a particular amino acid index, respectively.

Normalising each amino acid index sets the mean of the index to zero and standard deviation to one, this keeps all the amino acid indices consistent and avoids indices with extreme small or large values dominating in the analysis.

$$\frac{E' = E - \bar{x}(E)}{\sigma(E)} \quad \textbf{Eq 5-1}$$

An example of normalisation is shown in Table 5-1, which contains both raw and normalised values for AAI 1 ANDN92010. The mean and standard deviation of the raw index value are 4.4175 and 0.262782, respectively, and after normalisation, the mean and standard deviation are zero and one, respectively.

5.4 Novel feature extraction methods based on amino acid indices

5.4.1 Hybrid computational method for the analysis of amino acid indices – method 1

Each of the 611 amino acid indices is a unique set of values but some of these amino acid indices are similar to each other in terms of numerical values and property definition. This section is concerned with developing a computational method to combine similar amino acid indices together and then to derive a new computationally generated index to represent the original set of clustered amino acid indices. To represent this relationship among current amino acids indices, hierarchal cluster analysis and PCA methods are used to develop a hybrid method.

Hierarchical clustering is a technique that links similarity patterns found in data by the distance between data points based on the Euclidean distance function, with the general idea to merge similar amino acid indices into similar clusters. Hierarchical clustering is applied to the whole amino acid indices dataset (611x20), where 611 equals the number indices and 20 is the number descriptors per index, using three types of clustering modes (single, complete and average linkage). Within each of these clustered sets of amino acid indices PCA is applied to reduce the dimensionality of the indices and extracts the first principal component. An example, after clustering, cluster1 contains amino acid index 1 and 17 (matrix size 2 x 20). PCA is then applied over cluster1 - *aai* x 20 where *aai* equals the number of amino acid indices in the cluster (in this example 2). PCA returns a transposed matrix of 20 x *aai*, where the first component i.e. column1 is extracted and then transposed again to vector to 1x20. This new

vector (1 x 20) will have the largest variance which represent how much of the original amino acid indices in cluster 1 (index 1 and 17) it covers, a value of 1.0 means 100% variability of the original data is represented. Figure 5-1 illustrates the flow process for hybrid computational method.

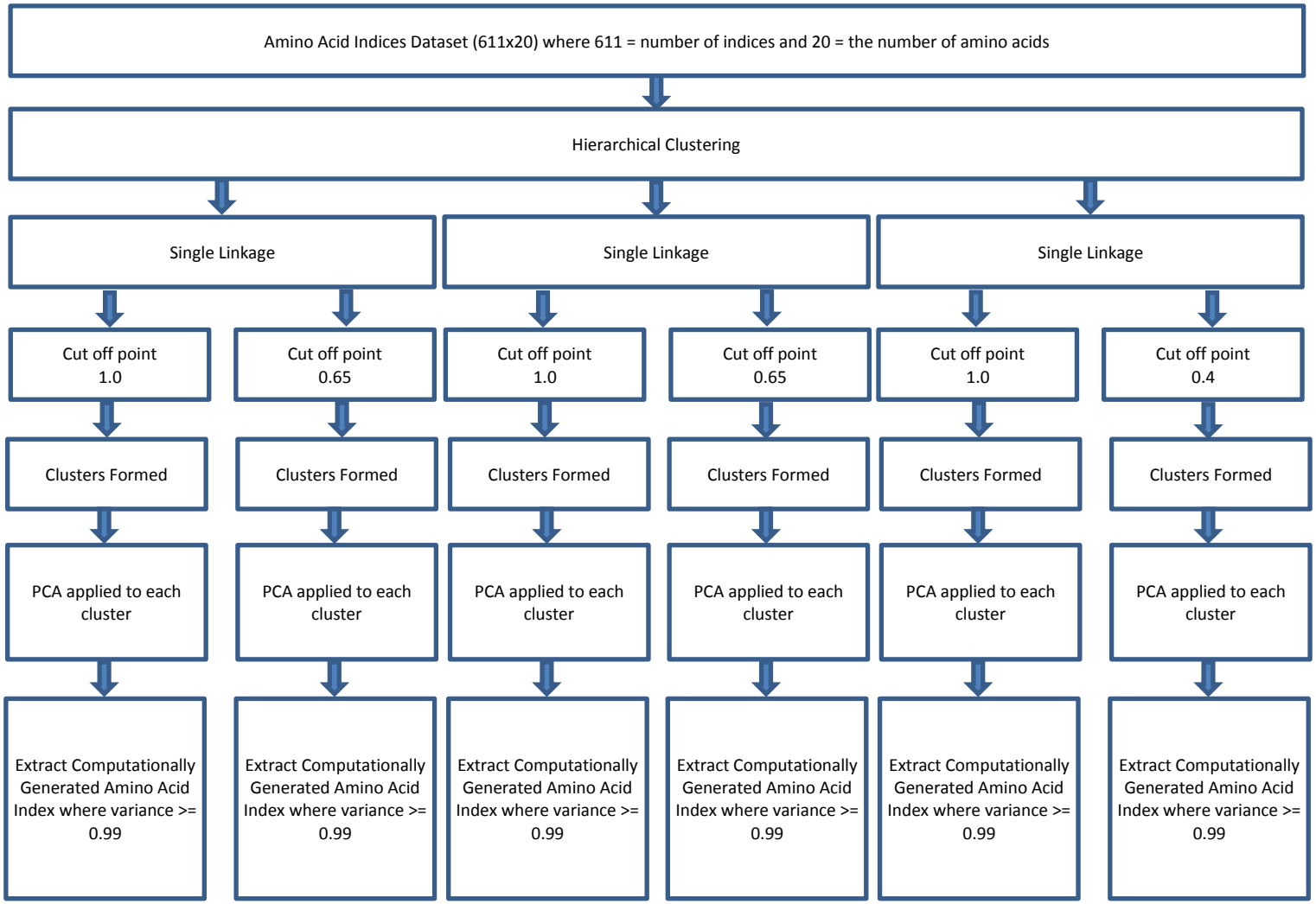


Figure 5-1 Hybrid Feature Reduction Method

5.4.2 Generalised amino acid composition – method 2

This section presents a novel sequence driven feature called generalised amino acid composition (GAAC). It is an adapted amino acid composition feature set where it gives a natural weight to each amino acid, in the form of an amino acid index.

In the literature, the amino acid composition feature group and as well as results presented in chapter 4 that it is a robust predictor with only twenty descriptor values. The amino acid composition feature group is a powerful and a widely used feature set that can be used to predict many proteomic characteristics such as protein protein-protein interaction (Roy, Martinez et al. 2009), protein function (Hansen, Lu et al. 2006) as well as protein structural class (Eisenhaber, Frömmel et al. 1996; Eisenhaber, Imperiale et al. 1996).

The results presented in Table 4-13 of chapter 4 the predictive accuracies achieved using the amino acid composition feature group ranges between 39.73%-57.88% across all the datasets and test procedure combinations. From the investigation carried out in chapter 4, it was found that two sequence-driven-features groups descriptor values are derived from amino acid indices, these are the autocorrelation and PseAAC sequence-driven feature groups.

Amino acid composition is based on the normalised frequency of the twenty amino acid types that appear in a proteins amino acid sequence (primary structure), however it does not take into consideration the natural weights of each amino acid type and instead places a weight of 1 to each amino acid. The values in the amino acid index can replace the weight of one for each amino acid type with its respective amino acid index value. With 611 amino acid indices (natural weights) available, there is potential to characterise a proteins predicted property by utilising an effective amino acid index using the GAAC method, thus translating into improved predictive accuracies. The generalisation aspect is the idea of having a general application use for the method; i.e. GAAC is not limited just to the prediction of protein structural classes or limited to a restricted set of amino acid indices, it can be generalised for many different properties and utilising all amino acid indices. The GAAC method is formulated in Eq 5-2.

$$GAAC(r) = \frac{n(r)w(r)}{\sum_{r=1}^{20} n(r)w(r)} \qquad \textbf{Eq 5-2}$$

where *GAAC(r)* is the generalised amino acid composition for amino acid *r*, *n(r)* is the frequency (or number) of the amino acid type *r* in a protein amino acid sequence and *w(r)* is the amino acid index value used as a weight for the amino acid type *r*.

5.4.3 Novel feature extraction methods over sequence representation matrix based on amino acid indices

The methods presented in this section are concerned with deriving new sequence driven features based a sequence representation matrix that utilises amino acid indices. The idea behind developing these methods is to represent an amino acid index from as single descriptor as opposed to a set of twenty descriptor values. Two methods are presented for the extraction of new sequence driven features using sequence representation matrix.

5.4.4 Sequence representation matrix

The first step to deriving new sequence driven features is to represent each protein sequence with all the amino acid indices as a single vector. The section presents two methods one using the average of all values (mean) and the other using principal component analysis (PCA). The first stage of the sequence representation is the same for both methods. An example, Table 5-1 contains the amino acid index values for index ANDN92010 and Figure 5-2 is a sample protein sequence. An example sequence representation matrix (SRM) process s to convert each amino acid residue in Figure 5-2 to its respective raw amino acid index value in Table 5-1. The sequence representation produces the resulting data in Table 5-2 This process will apply to each amino acid index (currently 611) to each protein sequence in any given protein dataset.

>1HTB:A| Homo sapiens (human) alcohol dehydrogenase
STAGKVIKCKAAVLWEVKKPFSIEDVEVAPPKAYEVRIKMVAVGICRTDDHVVSGNLVTPLPVILGHEAAGIVESVGEGVTTVKPGDKVIPLFTP
QCGKCRVCKNPESNYCLKNDLGNPRGTLQDGTRRFTCRGKPIHHFLGTSTFSQYTVVDENAVAKIDAASPLEKVCLIGCGFSTGYGSAVNVAKV
TPGSTCAVFGLGGVGLSAVMGCKAAGAARIIAVDINKDKFAKAKELGATECINPQDYKKPIQEVLKEMTDGGVDFSFEVIGRLDTMMASLLCC
HEACGTSVIVGVPPASQNLSINPMLLLTGRTWKGAVYGGFKSKEGIPKLVADFMAKKFSLDALITHVLPFEKINEGFDLLHSGKSICTVLTF

Figure 5-2 Protein Sequence for 1HTB taken from the Protein Data bank

Table 5-2 Sequence representation matrix, Result between Figure 5-2 protein sequence and Table 5-1 raw values

Sequence		Residue	Sequence		Residue	Sequence		Residue	Sequence		Residue	Sequence		Residue
1	S	4.5	56	N	4.75	111	C	4.65	167	E	4.29	222	V	3.95
2	T	4.35	57	L	4.17	112	L	4.17	168	K	4.36	223	D	4.76
3	A	4.35	58	V	3.95	113	K	4.36	169	V	3.95	224	I	3.95
4	G	3.97	59	T	4.35	114	N	4.75	170	C	4.65	225	N	4.75
5	K	4.36	60	P	4.44	115	D	4.76	171	L	4.17	226	K	4.36
6	V	3.95	61	L	4.17	116	L	4.17	172	I	3.95	227	D	4.76
7	I	3.95	62	P	4.44	117	G	3.97	173	G	3.97	228	K	4.36
8	K	4.36	63	V	3.95	118	N	4.75	174	C	4.65	229	F	4.29
9	C	4.65	64	I	3.95	119	P	4.44	175	G	3.97	230	A	4.35
10	K	4.36	65	L	4.17	120	R	4.38	176	F	4.29	231	K	4.36
11	A	4.35	66	G	3.97	121	G	3.97	177	S	4.5	232	A	4.35
12	A	4.35	67	H	4.63	122	T	4.35	178	T	4.35	233	K	4.36
13	V	3.95	68	E	4.29	123	L	4.17	179	G	3.97	234	E	4.29
14	L	4.17	69	A	4.35	124	Q	4.37	180	Y	4.6	235	L	4.17
15	W	4.7	70	A	4.35	125	D	4.76	181	G	3.97	236	G	3.97
16	E	4.29	71	G	3.97	126	G	3.97	182	S	4.5	237	A	4.35
17	V	3.95	72	I	3.95	127	T	4.35	183	A	4.35	238	T	4.35
18	K	4.36	73	V	3.95	128	R	4.38	184	V	3.95	239	E	4.29
19	K	4.36	74	E	4.29	129	R	4.38	185	N	4.75	240	C	4.65
20	P	4.44	75	S	4.5	130	F	4.29	186	V	3.95	241	I	3.95
21	F	4.29	76	V	3.95	131	T	4.35	187	A	4.35	242	N	4.75
22	S	4.5	77	G	3.97	132	C	4.65	188	K	4.36	243	P	4.44
23	I	3.95	78	E	4.29	133	R	4.38	189	V	3.95	244	Q	4.37
24	E	4.29	79	G	3.97	134	G	3.97	190	T	4.35	245	D	4.76
25	D	4.76	80	V	3.95	135	K	4.36	191	P	4.44	246	Y	4.6
26	V	3.95	81	T	4.35	136	P	4.44	192	G	3.97	247	K	4.36
27	E	4.29	82	T	4.35	137	I	3.95	193	S	4.5	248	K	4.36
28	V	3.95	83	V	3.95	138	H	4.63	194	T	4.35	249	P	4.44
29	A	4.35	84	K	4.36	139	H	4.63	195	C	4.65	250	I	3.95
30	P	4.44	85	P	4.44	140	F	4.29	196	A	4.35	251	Q	4.37
31	P	4.44	86	G	3.97	141	L	4.17	197	V	3.95	252	E	4.29
32	K	4.36	87	D	4.76	142	G	3.97	198	F	4.29	253	V	3.95
33	A	4.35	88	K	4.36	143	T	4.35	199	G	3.97	254	L	4.17
34	Y	4.6	89	V	3.95	144	S	4.5	200	L	4.17	255	K	4.36
35	E	4.29	90	I	3.95	145	T	4.35	201	G	3.97	256	E	4.29
36	V	3.95	91	P	4.44	146	F	4.29	202	G	3.97	257	M	4.52
37	R	4.38	92	L	4.17	147	S	4.5	203	V	3.95	258	T	4.35
38	I	3.95	93	F	4.29	148	Q	4.37	204	G	3.97	259	D	4.76
39	K	4.36	94	T	4.35	149	Y	4.6	205	L	4.17	260	G	3.97
40	M	4.52	95	P	4.44	150	T	4.35	206	S	4.5	261	G	3.97
41	V	3.95	96	Q	4.37	151	V	3.95	207	A	4.35	262	V	3.95
42	A	4.35	97	C	4.65	152	V	3.95	208	V	3.95	263	D	4.76
43	V	3.95	98	G	3.97	153	D	4.76	209	M	4.52	264	F	4.29
44	G	3.97	99	K	4.36	154	E	4.29	210	G	3.97	265	S	4.5
45	I	3.95	100	C	4.65	155	N	4.75	211	C	4.65	266	F	4.29
46	C	4.65	101	R	4.38	156	A	4.35	212	K	4.36	267	E	4.29
47	R	4.38	102	V	3.95	157	V	3.95	213	A	4.35	268	V	3.95
48	T	4.35	103	C	4.65	158	A	4.35	214	A	4.35	269	I	3.95
49	D	4.76	104	K	4.36	159	K	4.36	215	G	3.97	270	G	3.97
50	D	4.76	105	N	4.75	160	I	3.95	216	A	4.35	271	R	4.38
51	H	4.63	106	P	4.44	161	D	4.76	217	A	4.35	272	L	4.17
52	V	3.95	107	E	4.29	162	A	4.35	218	R	4.38	273	D	4.76
53	V	3.95	108	S	4.5	163	A	4.35	219	I	3.95	274	T	4.35
54	S	4.5	109	N	4.75	164	S	4.5	220	I	3.95	275	M	4.52
55	G	3.97	110	Y	4.6	165	P	4.44	221	A	4.35	276	M	4.52

277	A	4.35	307	L	4.17	337	A	4.35	367	S	4.5
278	S	4.5	308	L	4.17	338	K	4.36	368	I	3.95
279	L	4.17	309	L	4.17	339	K	4.36	369	C	4.65
280	L	4.17	310	T	4.35	340	F	4.29	370	T	4.35
281	C	4.65	311	G	3.97	341	S	4.5	371	V	3.95
282	C	4.65	312	R	4.38	342	L	4.17	372	L	4.17
283	H	4.63	313	T	4.35	343	D	4.76	373	T	4.35
284	E	4.29	314	W	4.7	344	A	4.35	374	F	4.29
285	A	4.35	315	K	4.36	345	L	4.17			
286	C	4.65	316	G	3.97	346	I	3.95			
287	G	3.97	317	A	4.35	347	T	4.35			
288	T	4.35	318	V	3.95	348	H	4.63			
289	S	4.5	319	Y	4.6	349	V	3.95			
290	V	3.95	320	G	3.97	350	L	4.17			
291	I	3.95	321	G	3.97	351	P	4.44			
292	V	3.95	322	F	4.29	352	F	4.29			
293	G	3.97	323	K	4.36	353	E	4.29			
294	V	3.95	324	S	4.5	354	K	4.36			
295	P	4.44	325	K	4.36	355	I	3.95			
296	P	4.44	326	E	4.29	356	N	4.75			
297	A	4.35	327	G	3.97	357	E	4.29			
298	S	4.5	328	I	3.95	358	G	3.97			
299	Q	4.37	329	P	4.44	359	F	4.29			
300	N	4.75	330	K	4.36	360	D	4.76			
301	L	4.17	331	L	4.17	361	L	4.17			
302	S	4.5	332	V	3.95	362	L	4.17			
303	I	3.95	333	A	4.35	363	H	4.63			
304	N	4.75	334	D	4.76	364	S	4.5			
305	P	4.44	335	F	4.29	365	G	3.97			
306	M	4.52	336	M	4.52	366	K	4.36			

5.4.5 Feature extraction using the mean of sequence representation matrix - method 3

The novel feature reduction method presented in this section is based on the mean; it will represent the average value of all amino acids and is computed as the sum of all the observed amino acid values from the protein sample divided by the length of the sequence of amino acids. The equation for the mean is defined in Eq 5-3 where $\bar{x}$ is the sample mean, N is the length of the protein sequence and the x corresponds to the observed value.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{N} \qquad \textbf{Eq 5-3}$$

Following on from the example given in the previous section 5.4.4, each amino acid residue of the protein amino acid sequence is converted to its respective amino acid index value, as shown in Table 5-2. The mean is then calculated from all the values presented in Table 5-2 (sum of all values is 221.12). The mean (4.3357) will form the sequence representation between the protein amino acid sequence in Figure 5-2 and the amino acid index Table 5-1. This effectively extracts a new feature from the sequence. The same procedure can be applied to all the amino acid indices, thus, converting a given protein amino acid sequence shown in Figure 5-2 into every amino acid index in the dataset (currently 611) will form a matrix of 611 x *N*, where *N* is the sequence length. The example conversion shown in Table 5-2 the matrix size is 611 x 374, as the sequence length is 374. After the mean extraction, the matrix size turned to 611x1. The same process is applied to each protein sequence in this study's four datasets, it will form the following matrix sizes, 25PDB (1168*x*611), 1189 (1085*x*611), Astral25 (5257*x*611) and Astral40 (7089*x*611), where the number on the left hands side of the *x* is the number of protein samples in that respective protein dataset.

5.4.6 Feature extraction using principal component analysis over sequence representation matrix - method 4

Method 4 is similar to method 3; it follows the same process up to the point shown in Table 5-2, where after instead of obtaining the mean from all the values, the method applies PCA. PCA will transform the 611xN (where 611 is the number of amino acid indices and *N* is the sequence length) into a set of principal components; the feature extraction method extracts the first principal component to form the sequence representation between the protein amino acid sequence in Figure 5-2 and the amino acid index in Table 5-1. The reason why the first principal component is extracted is that it will contain the largest variance out of all 611

principal components. Similarly with method 3, the same process is applied to each sequence in this study's four datasets, it will form the following matrix sizes, 25PDB (1168x611), 1189 (1085x611), Astral25 (5257x611) and Astral40 (7089x611). Where the number on the left hands side of the x is the number of protein samples in that respective protein dataset. Figure 5-3 illustrates the novel feature reduction method process for the mean and PCA for a single protein sample.

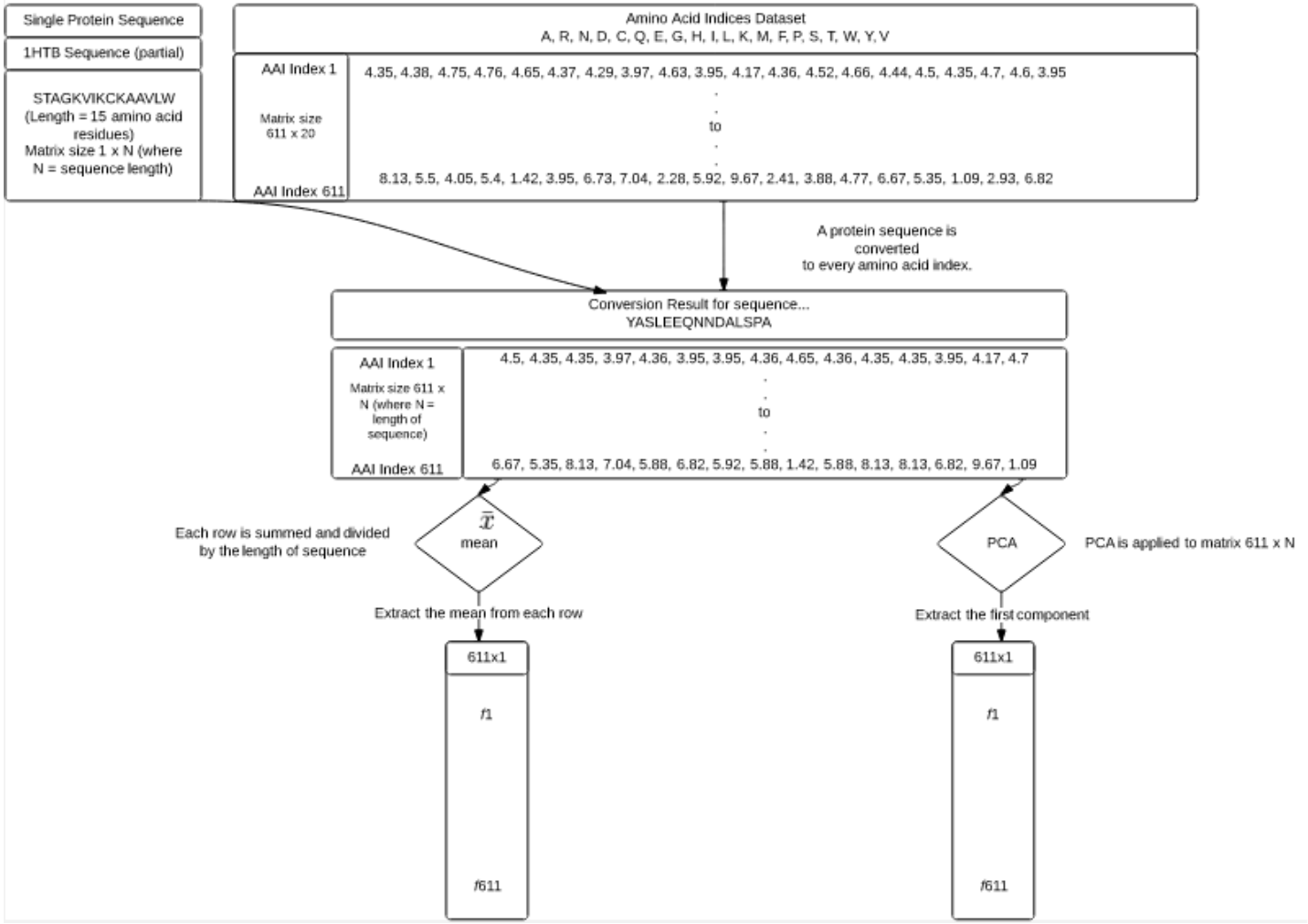


Figure 5-3 Flow diagram illustrating feature extraction

5.5 Results

A comprehensive analysis was undertaken to see how well each amino acid index predicts structural classes of protein. The main analysis consisted of classifying four-protein datasets each representing the full set of the amino acid indices evaluated against three test procedures (10-fold, leave-one-out an independent sets) using the mknn algorithm. Tens of thousands of results were obtained from the many hundreds of analyses undertaken. This resulted in 7332 individual sets of results, over twelve sets of dataset and test procedure combinations. The results shown in the subsequent tables are based on the highest achieved accuracies, which in all cases were obtained using the MKNN classifier. In an overall view, this study improves prediction results presented in chapter 4, which this study's results will be compared.

Table 5-3 Feature extraction method names

Full name of method	Short name of method
Hybrid based feature extraction	method 1
Generalised amino acid composition	method 2
Mean based feature extraction	method 3
PCA based feature extraction	method 4

5.5.1 Hybrid computational method for the analysis of amino acid indices reveals novel indices – Method 1

The idea behind the hybrid method was to see if there was any redundancy within the amino acid indices dataset and to remove them by clustering similar indices and then applying PCA over these clusters to extract a new amino acid index that represents >0.99% variability of the original cluster of similar amino acid indices. Nakai et al. (Nakai, Kidera et al. 1988) collected 222 amino acid indices from published literature and investigated the relationships among them using hierarchical cluster analysis, the same concept was taking to cluster the latest amino acid indices dataset. The first step of the hybrid method took into consideration clustering all the amino acid indices using three different linkage criteria's – single, complete and average. The cut off points was 1.0 and 0.65 for each linkage type, if the cut-off point was any higher, it resulted in all the indices being clustered in to a single cluster and if the cut-off point was any lower, each amino acid index is clustered on its own. The six hierarchal clustering methods and its number of generated clusters are shown in Table 5-11.

Table 5-4 Number of clusters generated

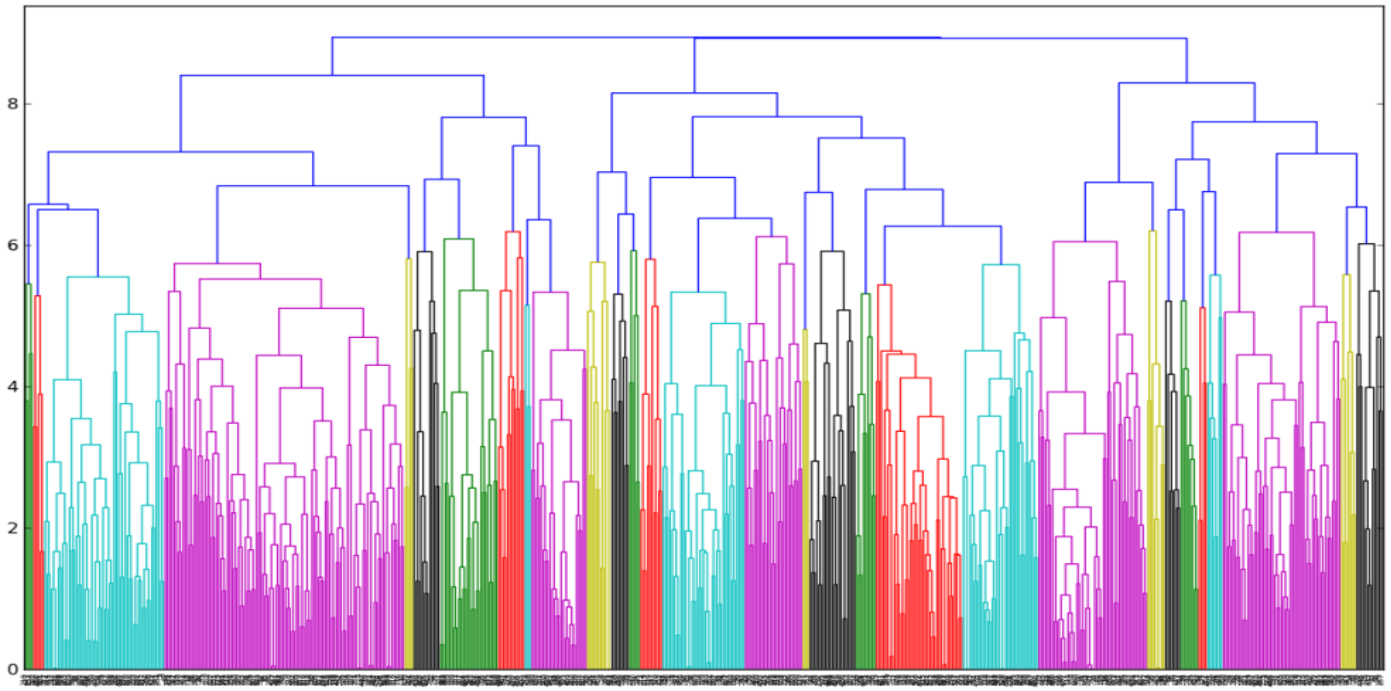
Hierarchical Clustering Methods	Cut-off point	No. Clusters Generated
Single Linkage	1.0	107
	0.65	134
Complete Linkage	1.0	181
	0.65	216
Average linkage	1.0	155
	0.4	155

Hierarchical clustering average linkage 1.0 and 0.4 methods produced the same number and arrangement of clusters. An example of a cluster is shown in Table 5-5 which is taken from hierarchical clustering single linkage method where minimum cluster distance = 1.0. The first cluster is formed by using amino acid index 1 and 17.

Table 5-5 Single Linkage and minimum Cluster Distance = 1, cluster 1 example

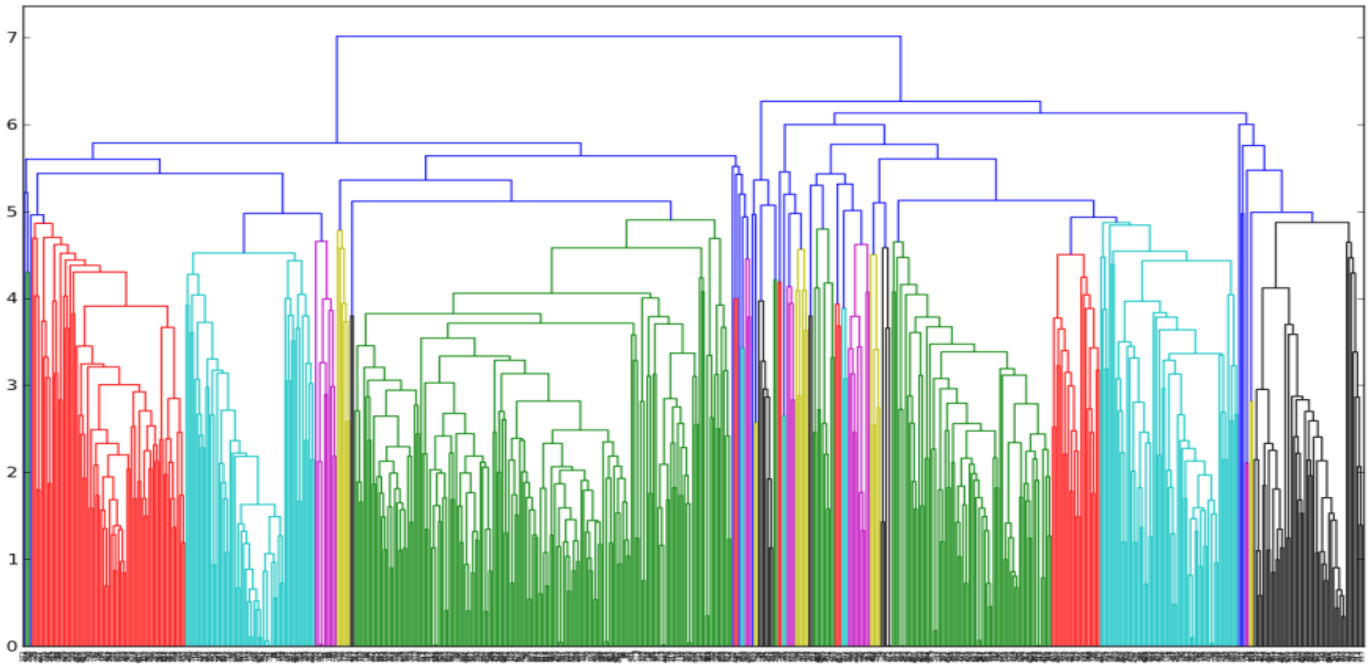
		Amino Acid Values									
Cluster 1 Index 1	**AAI # (1)**	**A**	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**
		4.35	4.38	4.75	4.76	4.65	4.37	4.29	3.97	4.63	3.95
		L	K	M	F	P	S	T	W	Y	V
		4.17	4.36	4.52	4.66	4.44	4.5	4.35	4.7	4.6	3.95
		Amino Acid Values									
Cluster 1 Index 2	**AAI # (17)**	**A**	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**
		4.34	4.39	4.75	4.76	4.68	4.37	4.29	3.97	4.6	4.22
		L	K	M	F	P	S	T	W	Y	V
		4.38	4.35	4.51	4.66	4.47	4.49	4.34	4.70	4.60	4.18

The arrangements within the clustered sets of amino acid indices were found to match the cross correlation values between indices. For example, where two or more indices are grouped together in a cluster, it means that the correlation between them is high, e.g. Index 1 and 17 appear in a cluster together using single linkage (where cut off point 1.0) - index 1, single linkage (cut off point 0.65) - index 1 and complete linkage (where cut off point 0.65) - index 1. Both indexes 1 and 17 relate to alpha-CH chemical shifts and have a correlation coefficient of 0.949, where a correlation coefficient close to 1 means they are highly correlated. Figure 5-4, Figure 5-5 and Figure 5-6 are the dendrogram produced by the hierarchical clustering method, which shows the distribution of amino acid indices and their dendrogram structures differ when analysed by means of the average, complete and single linkage based-hierarchical clustering methods, respectively. Hence, the different results obtained through the three hierarchical clustering methods.



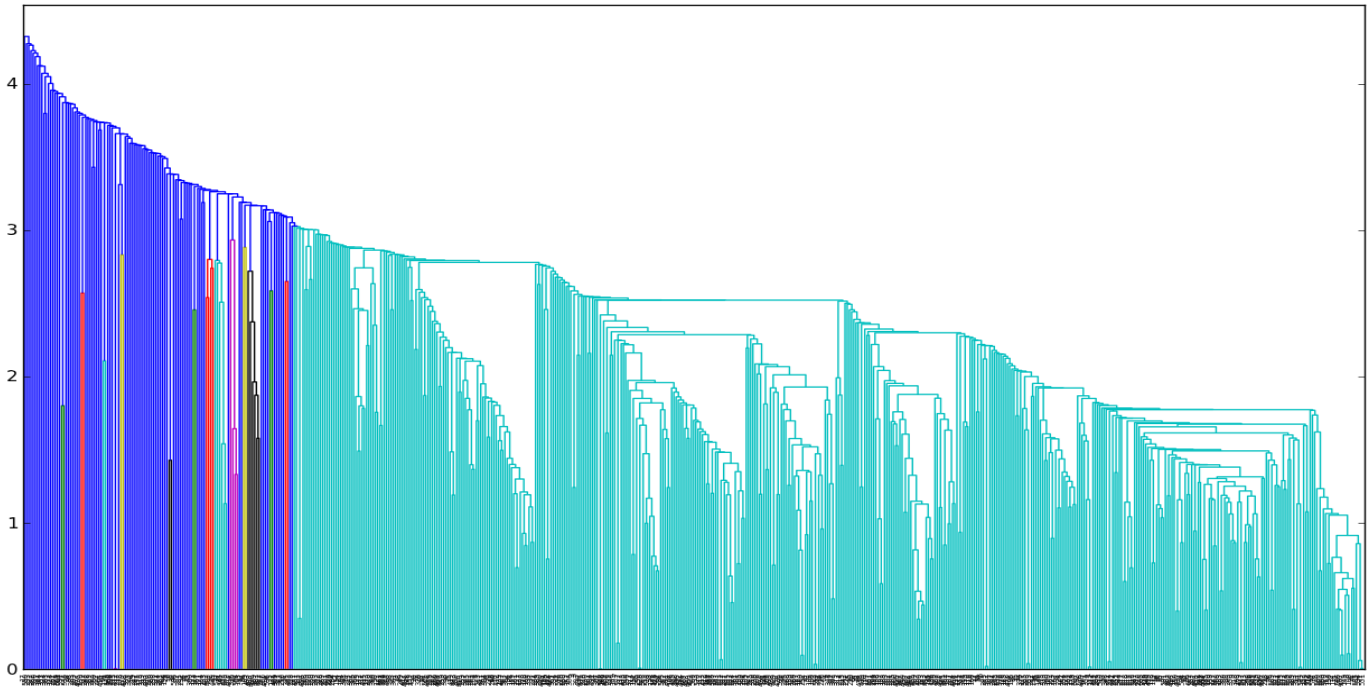
Amino Acid Indices listed in Appendix II

Figure 5-4 Average of Amino Acid Indices by using Complete Linkage based hierarchal clustering



Amino Acid Indices listed in Appendix II

Figure 5-5 Clustering of Amino Acid Indices by using Complete Linkage based hierarchal clustering



Amino Acid Indices listed in Appendix II

Figure 5-6 Clustering of Amino Acid Indices by using Single Linkage based on hierarchal clustering

PCA was then applied to each cluster within each of the five generated amino acid indices datasets. Computationally generated amino acid indices are then derived from the variances of PCA to summarise the amino acid indices within the cluster. Where clusters variance is ≥ 0.99% it was kept anything less discarded. The results in Table 5-6 show the number of new computationally generated amino acid indices with variance ≥ 0.99% and ≤ 0.99% variance. Where the variance is ≥ 0.99% describes that new computationally generated amino acid index produced by PCA represents 99% of the original clustered data variability. Where the computationally generated indices contain ≤ 0.99 variance, they are removed.

Table 5-6 Number of computationally derived indices

Hierarchical Clustering Methods	Cut-off point	No. of computationally generated indices ≥ 0.99 variance	No. of computationally generated indices ≤ 0.99 variance
Single Linkage	1.0	64	43
	0.65	133	1
Complete Linkage	1.0	81	100
	0.65	216	0
Average linkage	1.0	63	92
	0.4	63	92
Total		620	328

The new computationally generated amino acid indices were analysed using the 25PDB and 1189 datasets evaluated using independent-sets where the training dataset is Astral25 and Astral40, respectively. This was used because the most robust results are achieved where the largest datasets are used as training and evaluated using independent-sets test procedures. Results using the new computationally generated amino acid indices are presented in Table 5-7 for the 25PDB dataset and Table 5-8 for the 1189 dataset. All the results obtained shows that each computationally generated amino acid index achieved a higher accuracy than the original indices of the respective cluster. The accuracies presented in column "Accuracy (%) of generated index" in both tables is higher than the results presented in the column "Accuracy (%) if of Individual Index". By clustering highly correlated indices into one and replacing them with a summarised index removes noisy indices and redundancy from the amino acid dataset.

Table 5-7 Prediction Results Using Computationally Generated Amino Acid Indices testing dataset is 25PDB and training dataset is Astral25 using independent-sets test procedures

Hierarchical Clustering Methods	Cut off point	Generated index ID	Accuracy (%) of generated index	Summarised Indices	Accuracy (%) of Individual Index As shown in Table 5-11
Single Linkage	**1.0**	48	75.52%	AURR980110 (409)	55.16%
				AURR980115 (414)	57.91%
Single Linkage	**0.65**	105	75.52%	AURR980110 (409)	55.16%
				AURR980115 (414)	57.91%
Complete Linkage	**1.0**	4	67.87%	CHAM830108 (31)	65.17%
				QIAN880129 (284)	56.48%
Complete Linkage	**0.65**	176	75.52%	AURR980110 (409)	55.16%
				AURR980115 (414	57.91%
Average Linkage	**1.0**	60	65.09%	TdSc (582)	58.39%
	0.4			TdS (585)	56.42%

Table 5-8 Prediction Results Using Computationally Generated Amino Acid Indices testing dataset is 1189 and training dataset is Astral40 using independent-sets test procedures

Hierarchical Clustering Methods	Cut off point	Generated index ID	Accuracy (%) of generated index	Summarised Indices	Accuracy (%) of Individual Index As shown in Table 5-11
Single Linkage	**1.0**	89	65.11%	MEEJ800102 (178)	58.99%
				ZIMJ680105 (399)	53.82%
				Rf (554)	39.17%
				ProtScale_15 (596)	28.47%
Single Linkage	**0.65**	92	63.65%	SWER830101 (362)	54.84%
				CORJ870102 (515)	30.05%
Complete Linkage	**1.0**	110	63.65%	CHOP780214 (51)	52.44%
				ISOY800105 (123)	53.37%
				TANS770105 (367)	56.60%
Complete Linkage	**0.65**	158	63.65%	SWER830101 (362)	54.84%
				CORJ870102 (515)	30.05%
Average Linkage	**1.0**	104	65.11%	MEEJ800102 (178)	58.99%
				ZIMJ680105 (399)	53.82%
	0.4			Rf (554)	39.17%
				ProtScale_15 (596)	28.47%

Where the testing dataset is 1189, single and average linkage clusters where cut off point equals 1.0 produced the same set of results, along with single and complete clusters where cut off point equals to 0.65 also produced the same set of results. Where the testing dataset is 25PDB resulted in single linkage cut of point 1.0, single linkage cut off point 0.65 and complete

linkage cut off point 0.65 producing the same result from the same computationally generated amino acid index. The computationally generated amino acid index produced the highest result for the 25PDB dataset at 75.52%; this new amino acid index combines the old amino acid indices 409 – which relates to normalised positional residue frequency at helix termini N5 (Aurora and Rose 1998) and 414 which relates to normalized positional residue frequency at helix termini C1 (Aurora and Rose 1998). Helix capping are specific patterns of hydrogen bonding and hydrophobic interactions found at or near the ends of helices in proteins (Aurora and Rose 1998) which are highly related to alpha-helix structural class found in protein structures. It has been found that helix capping can play an important role in protein conformation; the capping process imposes a restriction on the number of possible conformations a protein structure can search through, which the same is reducing the conformation search space to find its tertiary structure (Aurora and Rose 1998; Eisenberg 2003). Which is why amino acid index 409 and 414 are picked up such analyses as the biological properties of such indices are related to protein structural classes, in particular amino acid index 414, which is a prominent amino acid index in feature extraction GAAC - method 2 results. The previous highest result produced from GAAC - method 2 produced 68.02% (listed in Table 5-11) using amino acid index number 414. The computationally generated amino acid indices have been shown to better represent individual indices by summarising them into a smaller set of amino acid indices, which generated revised dataset that is less noisy and non-redundant. Table 5-9 and Table 5-10 list the five new computationally generated amino acid indices that resulted in better predictive accuracy than its individual indices for the 25PDB and 1189 datasets, respectively. A full list of new computationally generated amino acid indices names and index numbers are presented in Appendices III to VII, index values available at the webserver http://cisaps.com/.

Table 5-9 Five best performing computationally generated indices for the 25PDB testing dataset trained using Astral25 dataset evaluated using independent-sets test procedure

Clustering Method	Single Linkage	Single Linkage	Complete Linkage	Complete Linkage	Average Linkage
Cut-Off Point	**1.0**	**0.65**	**1.0**	**0.65**	**1.0 and 0.4**
Index Name (AAI number – refer to appendix II and III for names)	AURR980110 (409)	AURR980110 (409)	CHAM830108 (31)	AURR980110 (409)	TdSc (582)
	AURR980115 (414)	AURR980115 (414)	QIAN880129 (284)	AURR980115 (414)	TdS (585)
New Generated Index ID (refer to appendices III to VIII for names)	48	105	4	176	60
(New) Amino Acid Values					
A	0.02905	0.15689	−0.0010	−0.0197	0.16238
R	0.02905	0.15689	−0.0010	−0.0197	0.16238
N	0.20849	0.18605	0.25127	0.07807	0.04475
D	0.02905	0.15689	−0.0010	−0.0197	0.16238
C	0.22963	−0.5111	0.02827	0.24041	0.41164
Q	0.02905	0.15689	−0.0010	−0.0197	0.16238
E	0.02905	0.15689	−0.0010	−0.0197	0.16238
G	0.20849	0.18605	0.25127	0.07807	0.04475
H	0.02905	0.15689	−0.0010	−0.0197	0.16238
I	0.22963	−0.5111	0.02827	0.24041	0.41164
L	0.02905	0.15689	−0.0010	−0.0197	0.16238
K	0.02905	0.15689	−0.0010	−0.0197	0.16238
M	0.20849	0.18605	0.25127	0.07807	0.04475
F	0.02905	0.15689	−0.0010	−0.0197	0.16238
P	0.22963	−0.5111	0.02827	0.24041	0.41164
S	0.02905	0.15689	−0.0010	−0.0197	0.16238
T	0.02905	0.15689	−0.0010	−0.0197	0.16238
W	0.20849	0.18605	0.25127	0.07807	0.04475
Y	0.02905	0.15689	−0.0010	−0.0197	0.16238
V	0.22963	−0.5111	0.02827	0.24041	0.41164

Table 5-10 Five best performing computationally generated indices for the 1189 testing dataset trained using Astral40 dataset evaluated using independent-sets test procedure

Clustering Method	**Single Linkage**	**Single Linkage**	**Complete Linkage**	**Complete Linkage**	**Average Linkage**
Cut-Off Point	**1.0**	**0.65**	**1.0**	**0.65**	**1.0 and 0.4**
Index Name (AAI number – refer to appendix II and III for names)	MEEJ800102 (178) ZIMJ680105 (399) Rf (554) ProtScale_15 (596)	SWER830101 (362) CORJ870102 (515)	CHOP780214 (51) ISOY800105 (123) TANS770105 (367)	SWER830101 (362) CORJ870102 (515)	MEEJ800102 (178) ZIMJ680105 (399) Rf (554) Proscale15 (596)
New Generated Index ID (refer to appendices III to VIII for names)	89	92	110	158	104
(New) Amino Acid Values					
A	−0.2379	0.23979	0.04854	0.23979	−0.2379
R	−0.1147	−0.1746	0.24901	−0.1746	−0.1147
N	0.22368	−0.1582	−0.1068	−0.1582	0.22368
D	0.17469	−0.1060	0.37057	−0.1060	0.17469
C	0.29202	0.16043	−0.3836	0.16043	0.29202
Q	0.01281	−0.1775	0.03059	−0.1775	0.01281
E	−0.2761	−0.1396	0.16501	−0.1396	−0.2761
G	−0.0832	0.22044	0.20217	0.22044	−0.0832
H	0.23289	−0.2583	−0.2700	−0.2583	0.23289
I	0.00497	0.33802	−0.0905	0.33802	0.00497
L	−0.2039	0.17557	0.01162	0.17557	−0.2039
K	0.06997	−0.0601	−0.0158	−0.0601	0.06997
M	−0.2151	0.06105	−0.0056	0.06105	−0.2151
F	0.08827	−0.2351	−0.3699	−0.2351	0.08827
P	0.25899	0.21331	0.35397	0.21331	0.25899
S	−0.2182	0.209	0.12315	0.209	−0.2182
T	0.05291	0.14821	0.13017	0.14821	0.05291
W	0.45192	−0.2567	−0.4310	−0.2567	0.45192
Y	−0.0667	−0.4734	0.06188	−0.4734	−0.0667
V	−0.4474	0.27358	−0.0735	0.27358	−0.4474

5.5.2 Assessment of amino acid indices for GAAC - method 2

All the features were extracted for each dataset using the GAAC method, which lead to in 7332 results. These are narrowed down to 120 to focus on the top results from each of the twelve combinations of dataset and test procedures analysis. The results presented in Table 5-11 are the top amino acid indices ranked for each dataset and test procedure analyses. Each dataset claims a different highest predicted amino acid index except between datasets 1189 and Astral25. Table 5-15 contains the top ranked results from each dataset and test procedure analysis of sequence-driven features.

Table 5-11 Highest predicted amino acid indices using each dataset

	10-fold			Leave-one-out			Independent-set		
	Index Name	**AAI No.**	**Accuracy**	**Index Name**	**AAI No.**	**Accuracy**	**Index Name**	**AAI No.**	**Accuracy**
25PDB	AURR980115	414	48.21%	AURR980115	414	48.38%	CHAM830108	31	65.17%
1189	KUMS000101	437	51.24%	KUMS000101	437	50.69%	CHAM830108	31	68.02%
Astral25	KUMS000101	437	47.90%	FUKS010109	466	49.52%	AURR980115	414	51.74%
Astral40	Nm	568	48.84%	DAYM780101	64	50.23%	Nm	568	49.44%

Index 31, 437 and 466 have a cross correlation > 0.8 between each other and between index 414 and 466 also has a cross correlation of > 0.8, value between +1 and 0 indicates a relationship between each index. Index 568 is a new index that was found through literature search that was not previously deposited in the original AAIndex database. The selection of amino acid indices shown in Table 5-11 are further discussed in chapter 7.

5.5.3 Comparison with the published and benched mark study results

Benched mark results are the results obtained analysing the largest set of sequence driven feature presented in Table 4-13 of chapter 4 and published results are the results presented in Table 2-8 which is a collection results of similar works that have been published by other authors.

The highest accuracies obtained for the 25PDB and 1189 datasets are 65.17% and 68.02%, respectively, using AAI CHAM830108 (index ID 31) see Table 5-14. Both results were obtained using independent-sets test procedure, which are the highest achieved using the generalised amino acid composition (GAAC) method. The importance of 25PDB and 1189 is that have been used in many studies and it has become a baseline measure to derive and compare results using these datasets. Compared to published results presented in Table 2-8 these results are higher using smaller feature size. The highest prediction accuracy for the two benched mark

datasets (25PDB and 1189) compared with the latest reported results are with the 25PDB dataset 64% (Mizianty and Kurgan 2009; Yang, Peng et al. 2009) and with the 1189 dataset 67.6% (Ke Chen 2008). The feature size varies with both sets of compared results (Mizianty and Kurgan 2009; Yang, Peng et al. 2009) paper uses a feature size of 160 and (Ke Chen 2008) uses a feature size of 50. While we do not claim to have the highest accuracy, we have obtained one of the highest accuracies in this field using the smallest and consistent feature set size of 20. The highest accuracy obtained for the Astral25 and Astral40 datasets are 51.74% and 50.23% using AAI AURR980115 (414) and AAI 64(DAYM780101), respectively.

GAAC method has improved protein structural class prediction accuracy over each highest predicted traditional sequence driven features presented in chapter 4 as shown in Table 5-12 to Table 5-14, values in bold shows GAAC achieved a better accuracy compared to the result presented in Table 4-13 chapter 4 results. The range of increase between chapter 4 and 5 results is between 4.15% to 9.82%

Table 5-12 Comparison of highest GAAC results and SDF using 10-fold test procedure

	10-fold				
	GAAC		Traditional SDF		Increase
	Chapter 5 - Table 5-15		Chapter 4 - Table 4-13		
Dataset	Accuracy %	AAI	Accuracy %	SDF	
25PDB	**48.21%**	414	41.69%	1	6.52%
1189	**51.24%**	437	41.42%	2	9.82%
Astral25	**47.90%**	437	41.40%	6.1	6.50%
Astral40	**48.84%**	568	40.33%	6.1	8.51%

Table 5-13 Comparison of highest GAAC results and SDF using leave-one-out test procedure

	Leave-one-out				
	GAAC		Traditional SDF		Increase
	Chapter 5 - Table 5-15		Chapter 4 - Table 4-13		
Dataset	Accuracy %	AAI	Accuracy %	SDF	
25PDB	**48.38%**	414	41.91%	1	6.47%
1189	**50.69%**	437	41.84%	2	8.85%
Astral25	**49.52%**	466	42.08%	6.1	7.44%
Astral40	**50.23%**	64	42.63%	1	7.60%

Table 5-14 Comparison of highest GAAC results and SDF using independent-sets test procedure

	Independent-sets				
	GAAC		Traditional SDF		Increase
	Chapter 5 - Table 5-15		Chapter 4 - Table 4-13		
Dataset	Accuracy %	AAI	Accuracy %	SDF	
25PDB	**65.17%**	31	60.79%	2	4.38%
1189	**68.02%**	31	63.87%	2	4.15%
Astral25	**51.74%**	414	44.30%	2	7.44%
Astral40	**49.44%**	568	41.38%	1	8.06%

5.5.3.1 Comparison of amino acid indices to traditional sequence driven feature utilising amino acid indices

Comparison against autocorrelation and PseAAC feature groups is important because they use a limited set of amino acid indices to derive the descriptor values for each of the feature groups, respectively it is an interest to see the effects of the full utilisation of amino acid indices. For example, autocorrelation feature group utilises eight amino acid indices (AAI 8, 9, 21, 22, 23, 33, 58 and 65) and PseAAC utilises three amino acid indices (AAI 534, 535 and 536). The comparison against amino acid composition is to highlight the effect of using amino acid indices as a natural amino acid weight.

The results presented in Table 5-16, Table 5-17, Table 5-18, Table 5-19 and Table 5-20 show that GAAC method improves over traditional sequence driven feature such as AAC, autocorrelation and PseAAC. GAAC results compared to traditional AAC and PseAAC, has increased between 0.00% - 10.01% and between 8.43% - 32.63%, respectively as shown in Table 5-16 across all four datasets. GAAC has also done better over autocorrelation feature groups as shown in Table 5-17 for 25PDB dataset, Table 5-18 for 1189 dataset, Table 5-19 for Astral25 dataset and Table 5-20 for Astral40 dataset where the GAAC method uses the same amino acid indices as the autocorrelation feature group (feature index 8, 9, 21, 22, 23, 33, 58 and 65).

Utilising amino acid indices to generalise the amino acid composition feature group has increased almost all results obtained in chapter 4. The ability to use any amino acid index to predict the protein structural class dataset is a flexible approach to identify which amino acid index is best suited out of the many hundreds that are available.

Table 5-15 Results obtained using GAAC– method 2 – refer to Table 5-3 for method names and Appendix II for AAI #.

Rank	AAI #	MKNN neighbours	25PDB	AAI #	MKNN neighbours	1189	AAI #	MKNN neighbours	Astral25	AAI #	MKNN neighbours	Astral40
10-fold												
1	414	[1,8,10,11]	48.21%	437	[3,4,5,8,10,11]	51.24%	437	[7,11]	47.90%	568	[1,7,9,10,11]	48.84%
2	456	[10,11]	45.67%	143	[1,4,8,9]	49.79%	568	[10,11]	47.82%	414	[8,11]	48.38%
3	495	[3,4,5,8,10,11]	45.67%	414	[10,11]	49.76%	466	[10,11]	47.76%	160	11	48.35%
4	343	[7,11]	45.61%	466	[5,8,9,11]	49.59%	346	[7,10]	47.71%	412	[9,11]	48.31%
5	198	[6,10]	45.45%	467	[1,3,5,6,9,11]	49.29%	414	[8,11]	47.69%	495	[10,11]	48.30%
6	160	[4,5,9,11]	45.38%	464	[6,11]	49.03%	64	[7,11]	47.66%	230	[7,9,10,11]	48.09%
7	408	[1,3,4,9,10,11]	45.32%	411	[8,10]	49.01%	409	[7,9,10,11]	47.58%	138	[3,9,10,11]	48.07%
8	610	[10,11]	45.27%	302	[4,6,7,9]	48.95%	230	[5,6,9,10]	47.56%	97	[10,11]	47.99%
9	467	[5,11]	45.23%	75	[3,4,7,11]	48.84%	137	[7,10]	47.45%	162	[4,8,9,10,11]	47.82%
10	568	[6,11]	45.19%	154	[1,3,8,11]	48.78%	464	[10,11]	47.45%	305	[5,7,10,11]	47.80%
Leave-one-out												
1	414	[10,11]	48.38%	437	[4,5,8,9]	50.69%	466	[3,6,9,11]	49.52%	64	[3,5,8,9,10,11]	50.23%
2	495	[7,11]	46.52%	467	[4,6,10,11]	50.69%	437	[9,11]	49.17%	437	[5,11]	50.22%
3	408	[7,11]	46.16%	411	[8,10]	50.05%	136	[8,11]	49.10%	532	[4,11]	50.22%
4	198	[4,11]	46.04%	466	[9,10]	50.05%	64	[7,11]	48.96%	134	[6,11]	50.12%
5	456	[7,11]	45.80%	143	[1,5,10,11]	49.86%	137	[9,11]	48.93%	468	[6,11]	50.09%
6	437	[4,5,9,10]	45.74%	414	[7,10]	49.86%	532	[7,10]	48.87%	136	[6,11]	50.04%
7	348	[5,10]	45.68%	464	[5,9,11]	49.77%	414	[10,11]	48.75%	414	[8,11]	50.01%
8	230	[5,10]	45.62%	463	[1,2,3,5,10,11]	49.49%	568	[10,11]	48.70%	31	[4,11]	49.97%
9	305	[8,11]	45.56%	302	[5,10]	49.40%	464	[9,11]	48.64%	568	[7,9,10,11]	49.94%
10	466	[3,7,9,11]	45.56%	31	[1,8,9,10,11]	49.22%	346	[7,10]	48.62%	456	[6,7,9,11]	49.92%
Independent-sets												
1	31	1	65.17%	31	1	68.02%	414	[1,2,5,8,10,11]	51.74%	568	[1,2,4,8,9,10,11]	49.44%
2	64	1	63.25%	28	1	66.64%	31	[1,11]	51.23%	414	[4,7,9,10]	49.20%
3	134	1	62.95%	437	1	66.64%	573	[1,2,5,7,9,11]	51.08%	138	11	48.08%
4	188	1	62.95%	201	1	66.36%	468	[1,2,4,7,9,11]	50.85%	170	[5,7,9,11]	47.96%
5	466	1	62.77%	137	1	66.08%	136	[1,2,4,7,9,10]	50.72%	119	[10,11]	47.83%
6	201	1	62.65%	196	1	65.99%	466	[1,2,5,8,9,11]	50.70%	412	[10,11]	47.80%
7	456	1	62.65%	302	1	65.90%	581	[1,2,6,8]	50.64%	599	[8,10]	47.80%
8	467	1	62.65%	64	1	65.81%	532	[1,2,5,9,10,11]	50.59%	437	[7,11]	47.77%
9	602	1	62.53%	453	1	65.71%	54	[1,2,6,9,10,11]	50.53%	98	[6,7,10,11]	47.68%
10	454	1	62.47%	457	1	65.53%	346	[1,6,10,11]	50.36%	32	[1,3,5,7,8,11]	47.52%

Table 5-16 Comparison of results obtained by AAC and PseAAC given in Table 4-13 (chapter 4) to those obtained in chapter 5 GAAC – method 2

	10-fold					Leave-one-out					Independent-sets				
	Table 4-13 Chapter 4		Table 5-11 Chapter 5		Increase	Table 4-13 Chapter 4		Table 5-11 Chapter 5		Increase	Table 4-13 Chapter 4		Table 5-11 Chapter 5		Increase
	Feature Index	Accuracy	AAI #	Accuracy		Feature Index	Accuracy	AAI #	Accuracy		Feature Index	Accuracy	AAI #	Accuracy	
25PDB	1	41.69%	414	48.21%	6.52%	1	41.91%	414	48.38%	6.47%	1	55.16%	31	65.17%	10.01%
	10	37.65%			10.56%	10	38.61%			9.77%	10	39.51%			25.66%
1189	1	41.00%	437	51.24%	10.24%	1	41.75%	437	50.69%	8.94%	1	57.88%	31	63.87%	5.99%
	10	36.79%			14.45%	10	38.62%			12.07%	10	42.67%			21.20%
Astral25	1	39.45%	437	47.90%	8.45%	1	41.56%	466	49.52%	7.96%	1	43.37%	414	44.30%	0.93%
	10	38.13%			9.77%	10	38.79%			10.73%	10	36.14%			8.16%
Astral40	1	39.73%	568	48.84%	9.11%	1	42.63%	64	50.23%	7.60%	1	41.38%	568	41.38%	0.00%
	10	39.78%			9.06%	10	38.96%			11.27%	10	35.52%			6.18%

Table 5-17 - Table 4-15 in Chapter 4 Autocorrelation feature group (its eight sub features are amino acid indices) comparison with Chapter 5 Amino Acid Indices that match the autocorrelation sub features for 25PDB dataset

25PDB																	
10-fold						Leave-one-out						Independent-sets					
Table 4-13 Chapter 4			Table 5-11 GAAC			Table 4-13 Chapter 4			Table 5-11 GAAC			Table 4-13 Chapter 4			Table 5-11 GAAC		
Feature Index	MKKN neighbours	%	AAI #	MKKN neighbours	%	Feature Index	MKKN neighbours	%	AAI #	MKKN neighbours	%	Feature Index	MKKN neighbours	%	AAI #	MKKN neighbours	%
Normalized Moreau-Borto Autocorrelation																	
3.1	[8,11]	33.03%	8	[7,11]	42.25%	3.1	10	33.21%	8	[6,11]	41.97%	3.1	1	54.44%	8	[1,2,3,7]	59.95%
3.2	[7,11]	31.58%	9	[10,11]	43.46%	3.2	[5,11]	32.01%	9	[6,10]	42.51%	3.2	1	49.04%	9	1	55.04%
3.3	[8,11]	28.90%	21	[7,10]	40.41%	3.3	11	29.98%	21	[8,10]	40.59%	3.3	1	47.78%	21	[1,2,4,5]	48.98%
3.4	[1,2,3,6,10,11]	27.80%	22	[4,10]	41.44%	3.4	[3,5,6,7]	27.88%	22	[4,11]	42.39%	3.4	1	44.90%	22	1	58.33%
3.5	[4,9,10,11]	29.25%	23	[4,7]	39.74%	3.5	[1,5,6,9,11]	28.30%	23	[4,8]	40.65%	3.5	1	49.52%	23	[1,2,3,4]	57.37%
3.6	[1,2,3,9,11]	29.92%	33	[6,11]	41.98%	3.6	[1,3,8]	29.62%	33	[7,11]	42.03%	3.6	1	48.14%	33	1	56.48%
3.7	[1,5,7,10,11]	30.95%	58	[6,10,11]	40.69%	3.7	[4,6,7,9]	31.84%	58	[5,6,10,11]	40.65%	3.7	1	51.20%	58	1	55.70%
3.8	[4,11]	28.67%	65	[4,11]	40.63%	3.8	[7,10]	28.00%	65	[4,11]	40.41%	3.8	1	47.12%	65	1	54.92%
Moran Autocorrelation																	
4.1	11	34.36%	8	[7,11]	42.25%	4.1	[5,9]	33.69%	8	[6,11]	41.97%	4.1	1	54.50%	8	[1,2,3,7]	59.95%
4.2	[5,11]	32.06%	9	[10,11]	43.46%	4.2	[1,2,3,4,6,8]	31.95%	9	[6,10]	42.51%	4.2	1	48.02%	9	1	55.04%
4.3	[5,6,9,11]	28.76%	21	[7,10]	40.41%	4.3	6	28.96%	21	[8,10]	40.59%	4.3	1	49.40%	21	[1,2,4,5]	48.98%
4.4	[1,3,5,7,9]	28.36%	22	[4,10]	41.44%	4.4	[1,2,5,6,7,10,11]	28.72%	22	[4,11]	42.39%	4.4	1	50.12%	22	1	58.33%
4.5	[8,10]	29.61%	23	[4,7]	39.74%	4.5	[4,8,10]	29.14%	23	[4,8]	40.65%	4.5	1	50.18%	23	[1,2,3,4]	57.37%
4.6	[5,6,8,9]	30.95%	33	[6,11]	41.98%	4.6	[1,5,6,8,9,10]	30.40%	33	[7,11]	42.03%	4.6	1	49.64%	33	1	56.48%
4.7	[1,2,6,7,9,10]	28.81%	58	[6,10,11]	40.69%	4.7	[6,9]	29.38%	58	[5,6,10,11]	40.65%	4.7	1	49.04%	58	1	55.70%
4.8	[1,3,5,9,10,11]	29.02%	65	[4,11]	40.63%	4.8	[3,5,6,9,11]	28.30%	65	[4,11]	40.41%	4.8	1	48.44%	65	1	54.92%
Geary autocorrelation																	
5.1	[4,8]	33.58%	8	[7,11]	42.25%	5.1	[3,4,5,9,10,11]	34.53%	8	[6,11]	41.97%	5.1	1	54.44%	8	[1,2,3,7]	59.95%
5.2	[9,11]	31.53%	9	[10,11]	43.46%	5.2	[4,7,10]	30.46%	9	[6,10]	42.51%	5.2	1	47.96%	9	1	55.04%
5.3	[9,11]	27.45%	21	[7,10]	40.41%	5.3	[1,7,8,10,11]	27.58%	21	[8,10]	40.59%	5.3	1	47.00%	21	[1,2,4,5]	48.98%
5.4	[1,3,7,9,11]	27.95%	22	[4,10]	41.44%	5.4	[1,2,5,9]	28.18%	22	[4,11]	42.39%	5.4	1	47.24%	22	1	58.33%
5.5	[9,11]	30.94%	23	[4,7]	39.74%	5.5	[5,6,8,9,10,11]	31.72%	23	[4,8]	40.65%	5.5	1	48.02%	23	[1,2,3,4]	57.37%
5.6	[7,10,11]	29.80%	33	[6,11]	41.98%	5.6	10	29.50%	33	[7,11]	42.03%	5.6	1	47.78%	33	1	56.48%
5.7	[5,11]	27.64%	58	[6,10,11]	40.69%	5.7	[5,11]	28.12%	58	[5,6,10,11]	40.65%	5.7	1	46.46%	58	1	55.70%
5.8	[3,5,6,9,10,11]	29.42%	65	[4,11]	40.63%	5.8	[3,6,9,10]	28.54%	65	[4,11]	40.41%	5.8	1	49.46%	65	1	54.92%

Table 5-18 Table 4 10 in Chapter 4 Autocorrelation feature group (its eight sub features are amino acid indices) comparison with Chapter 5 Amino Acid Indices that match the autocorrelation sub features for 1189 dataset

1189																		
10-fold						Leave-one-out						Independent-sets						
Table 4-13 Chapter 4			Table 5-11 GAAC			Table 4-13 Chapter 4			Table 5-11 GAAC			Table 4-13 Chapter 4			Table 5-11 GAAC		
Feature Index	MKNN neighbours	%	AAI #	MKNN neighbours	%	Feature Index	MKNN neighbours	%	AAI #	MKNN neighbours	%	Feature Index	MKNN neighbours	%	AAI #	MKNN neighbours	%
Normalized Moreau-Borto Autocorrelation																		
3.1	[3,4,9,10]	31.51%	8	[4,9]	46.83%	3.1	[9,10]	31.89%	8	[5,10]	46.82%	3.1	1	53.83%	8	[1,2,4,11]	62.95%
3.2	[1,8,9,11]	32.02%	9	[7,10]	47.10%	3.2	[1,8,10,11]	33.00%	9	[3,11]	47.19%	3.2	[1,2,4,10]	48.85%	9	[1,2,4,6]	58.71%
3.3	[4,5,11]	30.59%	21	[5,10]	43.86%	3.3	[3,4,5,6,7]	30.60%	21	[5,11]	44.06%	3.3	1	49.86%	21	[1,3]	51.89%
3.4	[1,6,8,11]	27.72%	22	[3,5,8,9]	46.20%	3.4	[3,5,7,9]	27.01%	22	[3,5,10,11]	46.73%	3.4	1	48.39%	22	1	62.12%
3.5	11	30.99%	23	[1,3,5,8,9,11]	44.99%	3.5	11	29.03%	23	[4,11]	45.81%	3.5	1	50.42%	23	1	59.54%
3.6	[1,9,10]	30.05%	33	[7,9,10,11]	47.11%	3.6	[3,9,10]	29.68%	33	[4,10]	48.20%	3.6	1	49.40%	33	[1,2,6,11]	59.54%
3.7	[1,3,5,10]	30.12%	58	[8,11]	44.74%	3.7	[1,2,3,6,7,8]	29.68%	58	[10,11]	45.53%	3.7	1	50.69%	58	1	58.34%
3.8	[1,3,6,7,8]	28.31%	65	[3,5,6,9,10,11]	43.96%	3.8	[3,6,8,9]	28.11%	65	[5,10]	44.98%	3.8	1	49.40%	65	1	60.00%
Moran Autocorrelation																		
4.1	[5,10,11]	32.37%	8	[4,9]	46.83%	4.1	[7,11]	32.63%	8	[5,10]	46.82%	4.1	1	55.12%	8	[1,2,4,11]	62.95%
4.2	[4,7,9]	33.03%	9	[7,10]	47.10%	4.2	[3,6,8,11]	33.55%	9	[3,11]	47.19%	4.2	1	51.61%	9	[1,2,4,6]	58.71%
4.3	[3,6,7,9,10]	33.24%	21	[5,10]	43.86%	4.3	[1,7,8,9,10,11]	33.00%	21	[5,11]	44.06%	4.3	1	51.34%	21	[1,3]	51.89%
4.4	[5,8,10,11]	28.95%	22	[3,5,8,9]	46.20%	4.4	[8,9]	28.39%	22	[3,5,10,11]	46.73%	4.4	1	49.49%	22	1	62.12%
4.5	[1,4,6,9]	30.78%	23	[1,3,5,8,9,11]	44.99%	4.5	[3,4,7,8]	29.77%	23	[4,11]	45.81%	4.5	1	52.90%	23	1	59.54%
4.6	[5,10,11]	30.95%	33	[7,9,10,11]	47.11%	4.6	[5,8,10,11]	30.78%	33	[4,10]	48.20%	4.6	1	51.15%	33	[1,2,6,11]	59.54%
4.7	[3,7,10,11]	28.95%	58	[8,11]	44.74%	4.7	[5,8,9,11]	29.22%	58	[10,11]	45.53%	4.7	1	51.89%	58	1	58.34%
4.8	11	27.84%	65	[3,5,6,9,10,11]	43.96%	4.8	[5,11]	27.83%	65	[5,10]	44.98%	4.8	1	48.66%	65	1	60.00%
Geary autocorrelation																		
5.1	[1,6,10,11]	32.81%	8	[4,9]	46.83%	5.1	[8,11]	31.80%	8	[5,10]	46.82%	5.1	1	54.10%	8	[1,2,4,11]	62.95%
5.2	[6,11]	32.02%	9	[7,10]	47.10%	5.2	[6,11]	32.26%	9	[3,11]	47.19%	5.2	1	51.89%	9	[1,2,4,6]	58.71%
5.3	[4,8]	32.51%	21	[5,10]	43.86%	5.3	[4,7]	31.89%	21	[5,11]	44.06%	5.3	1	51.52%	21	[1,3]	51.89%
5.4	[6,11]	30.23%	22	[3,5,8,9]	46.20%	5.4	[6,10]	29.77%	22	[3,5,10,11]	46.73%	5.4	1	47.83%	22	1	62.12%
5.5	[9,10]	30.40%	23	[1,3,5,8,9,11]	44.99%	5.5	[9,11]	31.71%	23	[4,11]	45.81%	5.5	[1,2,6,9]	50.60%	23	1	59.54%
5.6	11	33.52%	33	[7,9,10,11]	47.11%	5.6	[5,9,10]	33.00%	33	[4,10]	48.20%	5.6	1	52.07%	33	[1,2,6,11]	59.54%
5.7	[3,5,7,8,9]	30.67%	58	[8,11]	44.74%	5.7	[1,2,3,7,8,9,11]	29.59%	58	[10,11]	45.53%	5.7	1	49.59%	58	1	58.34%
5.8	[7,11]	28.01%	65	[3,5,6,9,10,11]	43.96%	5.8	[3,5,7,8,10]	28.11%	65	[5,10]	44.98%	5.8	1	49.12%	65	1	60.00%

Table 5-19 Table 4-15 in Chapter 4 Autocorrelation feature group (its eight sub features are amino acid indices) comparison with Chapter 5 Amino Acid Indices that match the autocorrelation sub features for Astral25 dataset

Astral25																	
10-fold						Leave-one-out						Independent-sets					
Table 4-13 Chapter 4			Table 5-11 GAAC			Table 4-13 Chapter 4			Table 5-11 GAAC			Table 4-13 Chapter 4			Table 5-11 GAAC		
Feature Index	MKNN neighbours	%	AAI#	MKNN neighbours	%	Feature Index	MKNN neighbours	%	AAI #	MKNN neighbours	%	Feature Index	MKNN neighbours	%	AAI#	MKNN K neighbours	%
Normalized Moreau-Borto Autocorrelation																	
3.1	[4,10,11]	33.25%	8	[10,11]	45.63%	3.1	[5,7,9,10,11]	34.05%	8	[5,11]	46.43%	3.1	[1,2,3,4,11]	38.62%	8	[1,2,3,5,7,8,10,11]	48.48%
3.2	[3,5,8,11]	30.82%	9	[8,11]	45.38%	3.2	[1,4,5,7,8,9]	31.16%	9	[8,11]	46.53%	3.2	1	35.33%	9	[1,2,4,8,10,11]	48.05%
3.3	[1,2,3,4,6,7,8]	27.24%	21	[4,7,9,10]	43.33%	3.3	[1,2,3,6,9,11]	27.54%	21	[5,9,10,11]	45.48%	3.3	1	33.61%	21	[1,2,7,8,10,11]	44.90%
3.4	[8,9,11]	27.94%	22	[3,9,11]	44.52%	3.4	[8,9,11]	27.68%	22	[3,6,8,9]	45.25%	3.4	1	34.83%	22	[1,2,3,7,10,11]	48.39%
3.5	9	28.65%	23	[7,11]	43.51%	3.5	[9,11]	29.33%	23	[5,6,9,11]	46.55%	3.5	1	35.72%	23	[1,11]	46.59%
3.6	[1,4,5,6,8,10,11]	27.54%	33	[4,9,10,11]	44.77%	3.6	[4,5,7,10,11]	28.23%	33	[8,11]	43.68%	3.6	1	33.26%	33	[1,2,3,7,10,11]	47.04%
3.7	[4,5,9]	28.59%	58	[4,9,10,11]	43.28%	3.7	5	28.63%	58	[9,10]	45.39%	3.7	1	34.57%	58	[1,2,8,10]	46.65%
3.8	[4,10,11]	28.44%	65	[7,11]	42.86%	3.8	[8,9,11]	28.15%	65	[7,10]	44.34%	3.8	1	34.76%	65	[1,2,4,7,9,10]	47.44%
Moran Autocorrelation																	
4.1	[1,4,5,9,10,11]	33.81%	8	[10,11]	45.63%	4.1	[9,10]	34.79%	8	[5,11]	46.43%	4.1	[1,2,7,8,11]	38.82%	8	[1,2,3,5,7,8,10,11]	48.48%
4.2	[3,8,9,11]	28.93%	9	[8,11]	45.38%	4.2	[8,11]	30.09%	9	[8,11]	46.53%	4.2	1	34.79%	9	[1,2,4,8,10,11]	48.05%
4.3	[8,9,10]	26.73%	21	[4,7,9,10]	43.33%	4.3	[7,9,10]	26.97%	21	[5,9,10,11]	45.48%	4.3	1	33.33%	21	[1,2,7,8,10,11]	44.90%
4.4	10	27.25%	22	[3,9,11]	44.52%	4.4	[7,10,11]	27.09%	22	[3,6,8,9]	45.25%	4.4	1	33.70%	22	[1,2,3,7,10,11]	48.39%
4.5	[5,8,10,11]	29.89%	23	[7,11]	43.51%	4.5	[8,10]	29.60%	23	[5,6,9,11]	46.55%	4.5	1	35.40%	23	[1,11]	46.59%
4.6	[7,9]	27.39%	33	[4,9,10,11]	44.77%	4.6	[3,5,9,10]	27.91%	33	[8,11]	43.68%	4.6	1	34.42%	33	[1,2,3,7,10,11]	47.04%
4.7	10	28.37%	58	[4,9,10,11]	43.28%	4.7	11	29.22%	58	[9,10]	45.39%	4.7	1	35.16%	58	[1,2,8,10]	46.65%
4.8	5	26.01%	65	[7,11]	42.86%	4.8	[1,6,8,9,10]	26.16%	65	[7,10]	44.34%	4.8	1	34.61%	65	[1,2,4,7,9,10]	47.44%
Geary autocorrelation																	
5.1	11	33.10%	8	[10,11]	45.63%	5.1	[8,11]	33.77%	8	[5,11]	46.43%	5.1	[1,2,4,8]	39.32%	8	[1,2,3,5,7,8,10,11]	48.48%
5.2	11	28.98%	9	[8,11]	45.38%	5.2	[7,9]	29.62%	9	[8,11]	46.53%	5.2	[1,2,3,4,10]	35.31%	9	[1,2,4,8,10,11]	48.05%
5.3	[1,6,7,8,10]	27.31%	21	[4,7,9,10]	43.33%	5.3	[1,3,8,9,10]	27.47%	21	[5,9,10,11]	45.48%	5.3	1	34.22%	21	[1,2,7,8,10,11]	44.90%
5.4	[1,3,9,10,11]	26.77%	22	[3,9,11]	44.52%	5.4	11	26.82%	22	[3,6,8,9]	45.25%	5.4	1	34.15%	22	[1,2,3,7,10,11]	48.39%
5.5	[4,6,10]	29.75%	23	[7,11]	43.51%	5.5	11	29.45%	23	[5,6,9,11]	46.55%	5.5	1	34.22%	23	[1,11]	46.59%
5.6	[8,9,10]	28.32%	33	[4,9,10,11]	44.77%	5.6	[1,6,8,9,10,11]	28.34%	33	[8,11]	43.68%	5.6	1	33.59%	33	[1,2,3,7,10,11]	47.04%
5.7	[5,7,9]	28.53%	58	[4,9,10,11]	43.28%	5.7	[5,7,11]	28.74%	58	[9,10]	45.39%	5.7	1	33.91%	58	[1,2,8,10]	46.65%
5.8	5	26.86%	65	[7,11]	42.86%	5.8	[4,5,6,8,10]	27.15%	65	[7,10]	44.34%	5.8	1	33.89%	65	[1,2,4,7,9,10]	47.44%

Table 5-20 Table 4-15 in Chapter 4 Autocorrelation feature group (its eight sub features are amino acid indices) comparison with Chapter 5 Amino Acid Indices that match the autocorrelation sub features for Astral40 dataset

Astral40																	
10-fold						Leave-one-out						Independent-sets					
Table 4-13 Chapter 4			Table 5-11 GAAC			Table 4-13 Chapter 4			Table 5-11 GAAC			Table 4-13 Chapter 4			Table 5-11 GAAC		
Feature Index	MKNN neighbours	%	AAI #	MKNN neighbours	%	Feature Index	MKNN neighbours	%	AAI #	MKNN neighbours	%	Feature Index	MKNN neighbours	%	AAI #	MKNN neighbours	%
Normalized Moreau-Borto Autocorrelation																	
3.1	[5,10,11]	34.20%	8	[9,11]	46.00%	3.1	[7,11]	35.13%	8	[6,10]	48.20%	3.1	[1,3,4,8,10,11]	34.67%	8	[3,11]	46.07%
3.2	[10,11]	29.31%	9	[9,11]	46.22%	3.2	[6,10,11]	31.40%	9	[4,11]	47.64%	3.2	[1,4,10,11]	32.49%	9	[4,11]	47.05%
3.3	[1,4,8,9,10]	29.05%	21	[10,11]	44.65%	3.3	[1,4,5,6,7,9,10]	28.42%	21	[8,11]	46.21%	3.3	1	30.65%	21	[3,5,10,11]	42.44%
3.4	[8,9,10,11]	27.27%	22	[7,10]	45.21%	3.4	10	28.06%	22	[6,11]	47.45%	3.4	1	29.88%	22	[10,11]	42.84%
3.5	[10,11]	28.78%	23	[7,10]	43.53%	3.5	[3,8,9,10,11]	29.41%	23	[8,10]	45.65%	3.5	1	31.66%	23	[1,4,7,11]	45.63%
3.6	[4,7,8,10,11]	28.44%	33	[6,7,10,11]	45.73%	3.6	[1,5,7,8,11]	27.63%	33	[4,5,7,9,10,11]	47.91%	3.6	1	30.46%	33	[8,10]	46.49%
3.7	[1,3,4,5,7,8,9]	28.67%	58	[10,11]	44.20%	3.7	[1,3,6,7,8,10,11]	29.13%	58	[8,10]	46.81%	3.7	1	29.97%	58	[1,6,8,10]	43.74%
3.8	11	27.26%	65	[8,11]	43.79%	3.8	[9,10,11]	0.27818	65	[5,11]	46.48%	3.8	[1,3,4,8,9,10]	0.2958	65	[1,3,4,8,9,11]	46.22%
Moran Autocorrelation																	
4.1	[3,4,5,9,10,11]	33.07%	8	[9,11]	46.00%	4.1	[10,11]	35.76%	8	[6,10]	48.20%	4.1	[1,7,8,11]	34.06%	8	[3,11]	46.07%
4.2	[5,9,11]	28.29%	9	[9,11]	46.22%	4.2	11	30.02%	9	[4,11]	47.64%	4.2	[1,2,3,4,5,9,10,11]	31.54%	9	[4,11]	47.05%
4.3	[5,8,9,10,11]	27.10%	21	[10,11]	44.65%	4.3	[4,8,10]	27.61%	21	[8,11]	46.21%	4.3	[1,3,5,6,11]	31.54%	21	[3,5,10,11]	42.44%
4.4	11	26.58%	22	[7,10]	45.21%	4.4	10	28.02%	22	[6,11]	47.45%	4.4	[1,2,3,10]	30.28%	22	[10,11]	42.84%
4.5	[3,7,11]	28.86%	23	[7,10]	43.53%	4.5	[10,11]	29.41%	23	[8,10]	45.65%	4.5	1	31.81%	23	[1,4,7,11]	45.63%
4.6	[4,10,11]	26.77%	33	[6,7,10,11]	45.73%	4.6	[9,10]	27.61%	33	[4,5,7,9,10,11]	47.91%	4.6	1	30.89%	33	[8,10]	46.49%
4.7	[6,8,11]	28.73%	58	[10,11]	44.20%	4.7	11	30.15%	58	[8,10]	46.81%	4.7	[1,3,4,8,9]	31.45%	58	[1,6,8,10]	43.74%
4.8	[3,4,11]	27.16%	65	[8,11]	43.79%	4.8	[8,10]	27.42%	65	[5,11]	46.48%	4.8	[1,2,3,4,6,7,8]	29.17%	65	[1,3,4,8,9,11]	46.22%
Geary autocorrelation																	
5.1	[4,5,8,10]	33.01%	8	[9,11]	46.00%	5.1	[4,6,8,10,11]	35.37%	8	[6,10]	48.20%	5.1	[1,2,3,4,10,11]	34.68%	8	[3,11]	46.07%
5.2	[6,8,10,11]	29.14%	9	[9,11]	46.22%	5.2	[4,5,6,10,11]	30.27%	9	[4,11]	47.64%	5.2	[1,2,3,4,7,8,9,11]	32.46%	9	[4,11]	47.05%
5.3	[1,2,4,5,7,8,9]	26.79%	21	[10,11]	44.65%	5.3	[9,10,11]	28.47%	21	[8,11]	46.21%	5.3	[1,2,3,4,10]	31.00%	21	[3,5,10,11]	42.44%
5.4	10	27.81%	22	[7,10]	45.21%	5.4	[1,3,6,7,8]	27.58%	22	[6,11]	47.45%	5.4	[1,2,3,4,11]	29.93%	22	[10,11]	42.84%
5.5	10	29.43%	23	[7,10]	43.53%	5.5	11	30.44%	23	[8,10]	45.65%	5.5	[1,2,4,11]	31.26%	23	[1,4,7,11]	45.63%
5.6	11	27.08%	33	[6,7,10,11]	45.73%	5.6	[4,8,9,10,11]	28.79%	33	[4,5,7,9,10,11]	47.91%	5.6	1	30.73%	33	[8,10]	46.49%
5.7	[4,6,7,8,9,10,11]	28.88%	58	[10,11]	44.20%	5.7	11	30.17%	58	[8,10]	46.81%	5.7	[1,2,3,5,6,8,10]	31.86%	58	[1,6,8,10]	43.74%
5.8	[1,2,5,6,7,10,11]	27.19%	65	[8,11]	43.79%	5.8	[3,5,6,9,11]	28.28%	65	[5,11]	46.48%	5.8	[1,2,3,4,7]	29.04%	65	[1,3,4,8,9,11]	46.22%

5.5.4 Individual class performance

The results presented in the previous sections are all based on overall accuracy across each structural class. The individual class performance shows the predictive accuracy for each class across each dataset. Table 5-21 and Table 5-22 list the highest predicted amino acid indices of each structural class in each dataset. The highest predicted amino acid index that represent each structural class in each dataset also appears in the top 10 amino acid indices listed in Table 5-15 i.e. the selection of amino acid indices are common. An example from the 1189 dataset and evaluated using independent-sets test procedure, for each structural class the following amino acid indices were ranked the highest index 31 for all-α, index 31 for all-β, index 28 for α/ β and index 457 for α+β. Except for index 37, indices 31, 28 and 457 all appear in the top 10 amino acid indices listed in Table 5-15. Each structural class across each dataset using all test procedures roughly claims a different highest predicted amino acid index; however, out of a possible 32 highest predicted amino indices there are 22 unique amino acid indices present in Table 5-21 and Table 5-22. This makes it very difficult to zone in and specify which amino acid index is capable of predicting a structural class. There are consistencies with the selection of amino acid indices, such as some of these indices are common for a certain structural class i.e. for both of the astral datasets, amino acid index 414 is the highest predicted index across four out of the six all-α structural class prediction using all three-test procedures. Amino acid index 31 appears as the highest predicted amino acid index across structural classes. Full set of individual class results are listed in appendix X.

Table 5-21 Individual class majority selected feature per dataset (25PDB/1189) /test procedure

Structural Class	Datasets											
	25PDB						1189					
	Test Procedures											
	10-fold		Leave-one-out		Independent-Sets		10-fold		Leave-one-out		Independent-sets	
	AAI #	Accuracy	AAI #	Accuracy	AAI #	Accuracy	AAI #	Accuracy	AAI #	Accuracy	AAI #	Accuracy
All-α	568	59.30%	198	64.25%	31	58.6%	464	51.16%	414	50.67%	31	66.37%
All-β	568	53.03%	408	59.41	31	62.59%	437	57.25%	414	56.85%	31	68.49%
α/β	456	65.96%	198	65.12%	466	65.99%	302	81.82%	31	87.58%	28	80.00%
α+β	495	26.50%	414	22.22%	31	61.68%	75	19.17%	463	19.17%	457	58.33%

Table 5-22 Individual class majority selected feature per dataset (Astral25 / Astral40) /test procedure

Structural Class	Datasets											
	Astral25						Astral40					
	Test Procedures											
	10-fold		Leave-one-out		Independent-Sets		10-fold		Leave-one-out		Independent-sets	
	AAI #	Accuracy	AAI #	Accuracy	AAI #	Accuracy	AAI #	Accuracy	AAI #	Accuracy	AAI #	Accuracy
All-α	230	**54.27%**	414	**55.12%**	414	**63.16%**	414	**52.74%**	414	**54.22%**	414	**58.50%**
All-β	568	**55.76%**	568	**56.45%**	414	**57.52%**	414	**54.38%**	414	**55.73%**	412	**57.44%**
α/β	466	**73.98%**	437	**79.92%**	54	**72.16%**	568	**70.86%**	31	**62.86%**	32	**81.01%**
α+β	230	**22.81%**	466	**19.87%**	31	**29.17%**	495	**20.97%**	456	**22.65%**	32	**16.57%**

5.5.5 Assessment of performance based on test procedures

Boxplot views of the range of results across the entire amino acid index space for each test procedure (10-fold, leave-one-out and independent-set) is shown Figure 5-8, Figure 5-9 and Figure 5-10, respectively. Similarly, to chapter 4, the results presented in chapter 5 GAAC method shows that each test procedure affects the accuracies differently across each dataset using the same set of amino acid indices and the selection of these indices.

10-fold and leave-one-out test procedures output similar set of results, with leave-one-out outputting a difference between -1.56% to 2.18% for the 25PDB dataset, -1.59% to 2.40% for the 1189 dataset, -1.96 to 5.59% for the Astral25 dataset and -1.76 to 4.12% for the Astral40 dataset. Compared to the independent sets test procedure where testing datasets are 25PDB and 1189, the difference increase from 10-fold is between -3.11% – 21.73% for the 25PDB dataset and is between -2.81% – 21.58% for the 1189 dataset. Compared to leave-one-out is between -2.220% to 21.88% and for the 25PDB dataset, is between -2.64% – 20.18% for the 1189 dataset. Results are more robust and achieve the highest accuracies by using independent-sets test procedure where the testing dataset are 25PDB and 1189 are trained on the larger Astral25 and Astral40 datasets, respectively. However, these results come at the expense at longer computational analysis times are needed to train the *mknn* classifier on the larger astral datasets.

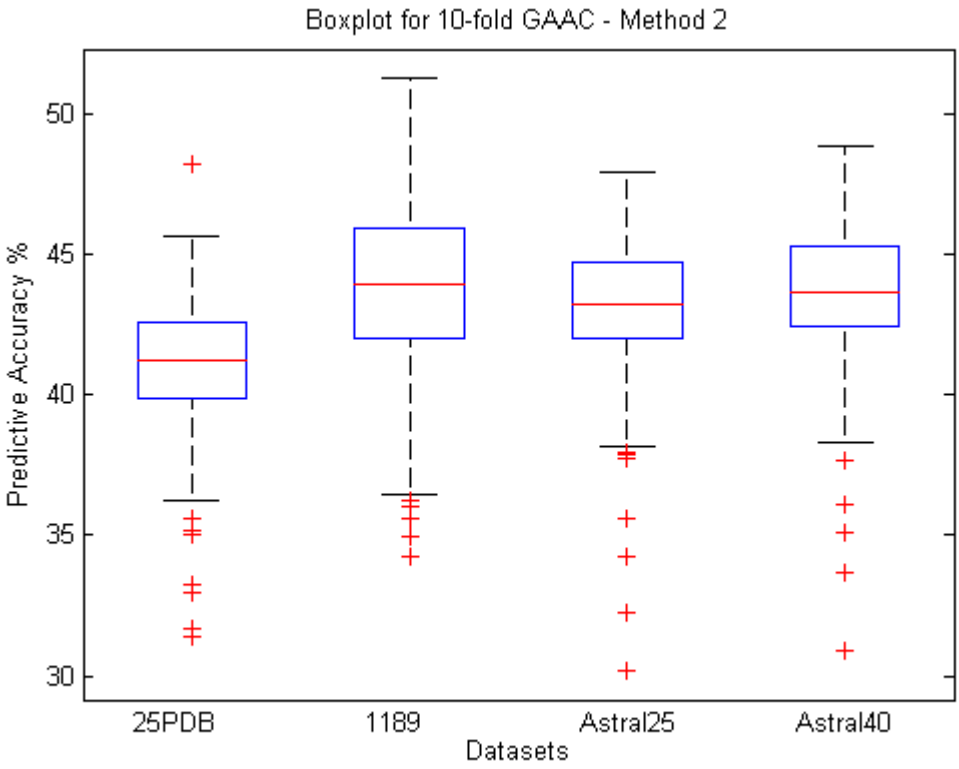


Figure 5-8 Boxplot for the GAAC using 10-fold test procedure

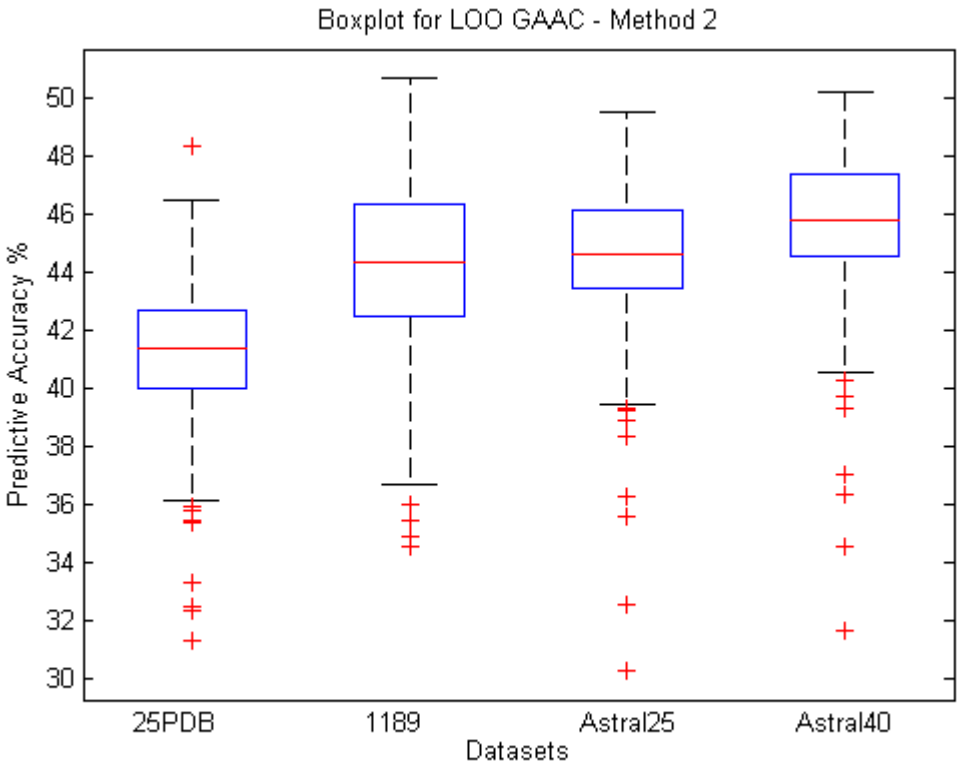


Figure 5-9 Boxplot for the GAAC using leave-one-out test procedure

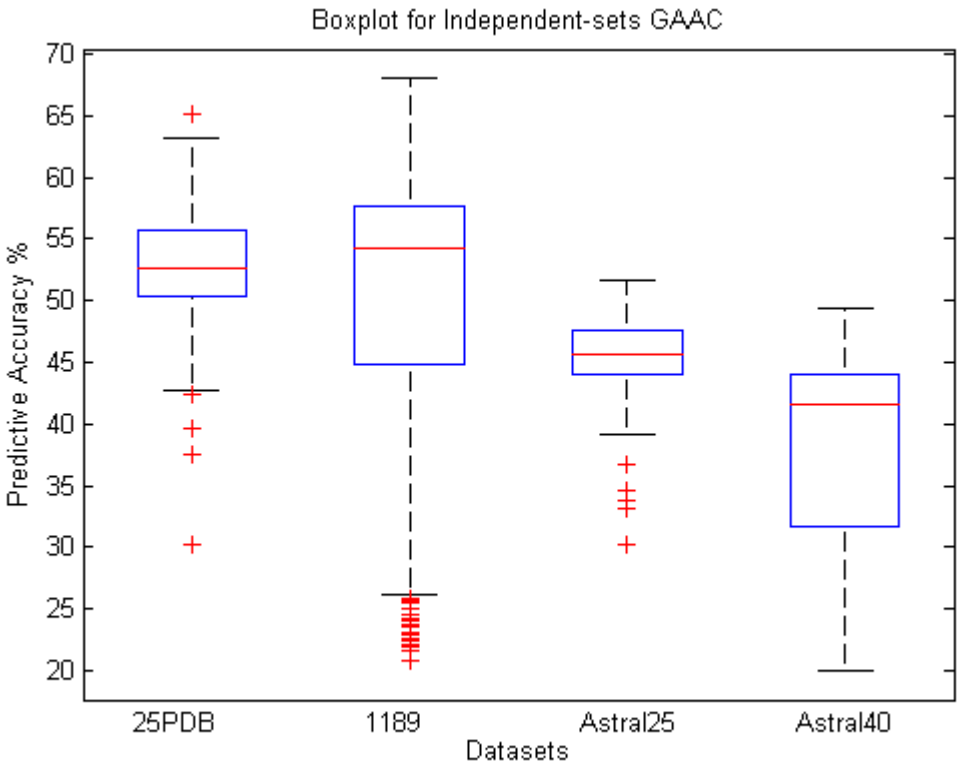


Figure 5-10 Boxplot for the GAAC using independent-sets test procedure

5.5.6 Results obtained using the novel feature extraction methods based on amino acid indices – methods 3 and 4

Two feature extraction methods were developed which were concerned with deriving new sequence driven feature based on the amino acid indices. The analysis carried out was also based on the same format as the GAAC method, which is to analyse both extracted feature sets across each of the four datasets evaluated using the three different test procedures. The final set of results did not achieve anything higher than what has been obtained through the GAAC method. Results are shown in Table 5-23 and Table 5-24 for the mean and PCA extraction methods, respectively. The principal idea behind extracting features using the mean and PCA was in theory a good idea however, the lower accuracy rate shows that a lot of information is lost through the extraction process. However, comparison between the mean and PCA derived sequence driven features, the sequence representation using the mean in all instances was marginally higher than PCA results. Method 4 - PCA results ranges between 28.2%-55.94% and the results obtained using the extraction method utilising the method 3 – mean are between 39.02% - 59.36%, whereas the GAAC method ranges between 45.19% - 69.02%.

Table 5-23 Results obtained using feature extraction method 3 – refer to Table 5-3 for method names and Appendix II for AAI #.

Rank	AAI #	MKNN neighbours	25PDB	AAI #	MKNN neighbours	1189	AAI #	MKNN neighbours	Astral25	AAI #	MKNN neighbours	Astral40
10-fold												
1	568	[9,11]	42.26%	412	[3,10,11]	45.80%	97	11	42.71%	568	11	43.19%
2	415	[3,8,9,10]	41.32%	222	[10,11]	44.13%	568	11	42.70%	222	[10,11]	42.12%
3	326	[10,11]	41.19%	568	[1,3,10,11]	43.67%	305	11	42.69%	230	11	41.80%
4	599	[4,9]	40.99%	408	[5,8]	43.14%	162	11	41.79%	305	11	41.70%
5	414	11	40.95%	414	[1,2,6,8,9,10,11]	42.67%	412	11	41.70%	414	11	41.69%
6	138	10	40.21%	411	[1,2,8,10,11]	42.38%	160	[1,5,6,9,10,11]	41.51%	97	11	41.67%
7	162	9	39.74%	261	[7,8,9,11]	42.38%	222	11	41.40%	265	11	41.43%
8	140	[3,8,10,11]	39.72%	230	10	42.31%	138	11	41.38%	412	[10,11]	41.40%
9	260	[9,11]	39.22%	100	[10,11]	42.29%	230	[3,5,9,10,11]	41.37%	140	11	41.25%
10	408	[1,5,10,11]	39.02%	410	[1,8,10,11]	41.91%	100	11	41.35%	138	[7,8,9,10,11]	41.20%
Leave-one-out												
1	568	[9,11]	41.97%	412	[4,5,10,11]	45.25%	97	11	42.95%	568	[9,11]	43.21%
2	326	[9,11]	41.25%	261	[1,9,10,11]	44.15%	568	[3,10,11]	42.67%	222	[7,9,10,11]	42.62%
3	414	[1,3,9,10]	41.19%	568	[1,2,4,8,9,11]	43.96%	305	[9,11]	42.59%	230	[3,8,10,11]	42.26%
4	415	[1,4,8,9,10]	40.71%	222	[10,11]	43.69%	162	11	42.00%	414	[6,7,9,10,11]	42.07%
5	162	[9,10,11]	40.35%	414	[8,10]	43.69%	230	[5,7,10,11]	41.83%	305	[5,9,10,11]	42.04%
6	599	[4,11]	40.35%	230	10	43.41%	265	[8,11]	41.77%	162	[3,5,7,9,10,11]	41.90%
7	100	[10,11]	40.11%	408	[5,8]	42.77%	412	[10,11]	41.66%	140	11	41.73%
8	260	9	39.87%	100	[7,11]	42.58%	222	11	41.64%	97	11	41.70%
9	254	[5,7,10,11]	39.81%	160	[7,8,10,11]	42.30%	160	[3,5,6,8,9,10,11]	41.58%	265	11	41.70%
10	91	[1,6,7,11]	39.69%	19	[4,8,9,11]	42.21%	254	[9,11]	41.53%	170	11	41.61%
Independent-sets												
1	412	1	56.60%	230	1	59.36%	414	[1, 2, 9, 11]	43.35%	160	11	44.01%
2	414	1	56.24%	223	1	58.80%	568	[1, 3, 4, 9, 10, 11]	42.71%	568	11	43.76%
3	140	1	56.18%	162	1	58.25%	305	[1, 2, 3, 6, 9, 10, 11]	42.52%	162	11	42.79%
4	415	1	56.06%	305	1	58.25%	412	[1, 2, 3, 7, 11]	42.34%	138	[1, 10, 11]	42.76%
5	91	1	55.94%	138	1	57.88%	411	[1, 2, 5, 6, 8, 10, 11]	42.23%	230	[1, 8, 10, 11]	42.76%
6	305	1	55.58%	599	1	57.79%	160	[1, 2, 3, 4, 7, 9, 10, 11]	42.19%	412	10	42.73%
7	38	1 2 3 9	55.40%	86	1	57.51%	265	[1, 6, 9, 11]	42.04%	414	11	42.57%
8	233	1	55.40%	339	1	57.51%	162	[1, 2, 5, 6, 10, 11]	41.89%	100	[1, 4, 9, 10, 11]	42.55%
9	331	1	55.28%	222	1	57.14%	230	[1, 5, 6, 10, 11]	41.81%	119	11	42.45%
10	599	1	55.28%	564	1	57.14%	415	[1, 2, 3, 10, 11]	41.53%	222	11	42.38%

Table 5-24 Results obtained using feature extraction method 4 – refer to Table 5-3 for method names and Appendix II for AAI #.

Rank	AAI #	MKNN neighbours	25PDB	AAI #	MKNN neighbours	1189	AAI #	MKNN neighbours	Astral25	AAI #	MKNN neighbours	Astral40
10-fold												
1	166	11	41.35%	45	[6,9,11]	42.79%	166	11	40.26%	81	11	40.64%
2	553	[5,9,11]	41.31%	553	[6,11]	41.95%	121	11	40.17%	581	[6,10,11]	39.69%
3	78	[6,8,9,11]	41.30%	396	[6,11]	41.66%	167	10	39.82%	128	[9,10,11]	39.56%
4	403	11	41.30%	167	[9,10]	41.48%	288	[9,10,11]	39.79%	521	11	39.29%
5	40	[1,4,9,10,11]	41.28%	78	11	41.39%	276	11	39.52%	578	[5,10,11]	39.21%
6	295	[9,10]	41.19%	51	[4,10,11]	41.10%	600	11	39.29%	600	11	39.08%
7	107	11	41.11%	37	11	40.91%	78	[3,4,9,10,11]	39.22%	170	11	39.07%
8	104	[3,4,7,9,10,11]	41.03%	166	[3,9,10,11]	40.58%	608	11	39.18%	238	11	39.01%
9	164	[7,10,11]	40.95%	403	[4,7,10,11]	40.53%	255	[5,10,11]	39.10%	479	11	38.98%
10	396	[1,7,9,10,11]	40.66%	333	[1,2,8,10,11]	40.45%	59	11	39.07%	80	11	38.95%
Leave-one-out												
1	295	11	41.91%	45	10	43.78%	166	[7,9,10,11]	41.20%	581	[8,10,11]	40.84%
2	166	11	41.61%	15	[3,8,9,11]	41.48%	121	11	40.69%	81	[9,10,11]	40.77%
3	255	[3,5,9,10,11]	41.49%	37	11	41.38%	608	11	40.44%	128	[3,8,9,10,11]	40.56%
4	104	[4,7,9,10]	41.31%	341	[5,7,8,10]	41.38%	167	10	40.40%	578	[6,8,9,10,11]	40.50%
5	553	[4,5,10,11]	41.25%	403	[10,11]	41.01%	288	11	40.38%	327	[9,10,11]	40.26%
6	396	11	41.13%	78	10	40.92%	7	[5,8,11]	40.19%	521	11	40.25%
7	107	[4,6,8,9,11]	41.01%	167	[7,10,11]	40.92%	253	[8,10,11]	40.10%	183	[7,10,11]	40.22%
8	40	[5,10,11]	40.89%	553	[5,11]	40.92%	287	[1,9,11]	39.99%	344	11	40.13%
9	78	[5,7,9,10]	40.83%	396	[5,11]	40.55%	600	[8,9,10]	39.95%	267	10	40.02%
10	37	[8,9,11]	40.77%	402	[9,10,11]	40.28%	82	[3,4,7,8,10,11]	39.89%	487	11	39.98%
Independent-sets												
1	428	1	55.94%	390	1	33.18%	7	[1, 3, 7, 9, 11]	42.72%	190	[9, 10]	29.09%
2	107	1	55.88%	385	1	32.54%	345	[1, 2, 3, 5, 7, 8, 9, 10, 11]	42.42%	597	[8, 10]	20.67%
3	82	1	55.82%	198	1	30.14%	377	[1, 2, 6, 7, 8, 10, 11]	42.36%	192	[8, 9, 10]	29.20%
4	274	1	55.70%	241	1	29.49%	166	[1, 2, 4, 6, 7, 9, 10]	42.34%	460	[7, 11]	29.93%
5	288	1	55.64%	445	1	29.22%	295	[1, 3, 6, 8, 9, 10, 11]	42.25%	397	[6, 9]	30.00%
6	290	1	55.40%	11	1	28.94%	600	[1, 2, 5, 6, 11]	42.23%	385	[6, 8, 10, 11]	36.27%
7	372	1	55.40%	444	1	28.85%	276	[1, 2, 4, 6, 7, 9, 11]	42.13%	439	[5, 9]	20.44%
8	35	1	55.22%	549	1	28.48%	287	[1, 3, 9, 10, 11]	42.13%	537	[4, 10]	19.88%
9	40	1	55.22%	446	1	28.30%	288	[1, 2, 4, 5, 7, 8, 9, 10, 11]	42.12%	445	[4, 5, 8, 9, 10]	28.48%
10	50	1	55.22%	397	1	28.20%	167	[1, 2, 7, 8, 11]	42.06%	448	[3, 9]	31.74%

5.6 Generalised Amino Acid Composition webserver

The positive results using GAAC resulted in the creation of a webserver where users can use GAAC to represent protein sequence datasets. The system is developed using PHP and MySQL. The front end graphical user interface is developed using PHP, which is a widely and highly configurable scripting language that is especially suited for Web development and can be embedded into HTML and the storage of data is stored using MySQL is a popular open-source database system. PHP combined with MySQL is a cross-platform web-developing tool that can be developed and viewed from any platform (i.e. Linux, Windows, Unix etc.) Figure 5-11 is a screen shot of the front end, it is the first screen a user sees when visiting the website, the user must enter an email address and input a protein sequence or upload a file containing many protein sequences in FASTA format to convert protein sequences. In bioinformatics, the FASTA format is a text format for representing protein sequences, in which the first line is the header information and subsequent lines are the amino acid residues represented using single-letter codes.

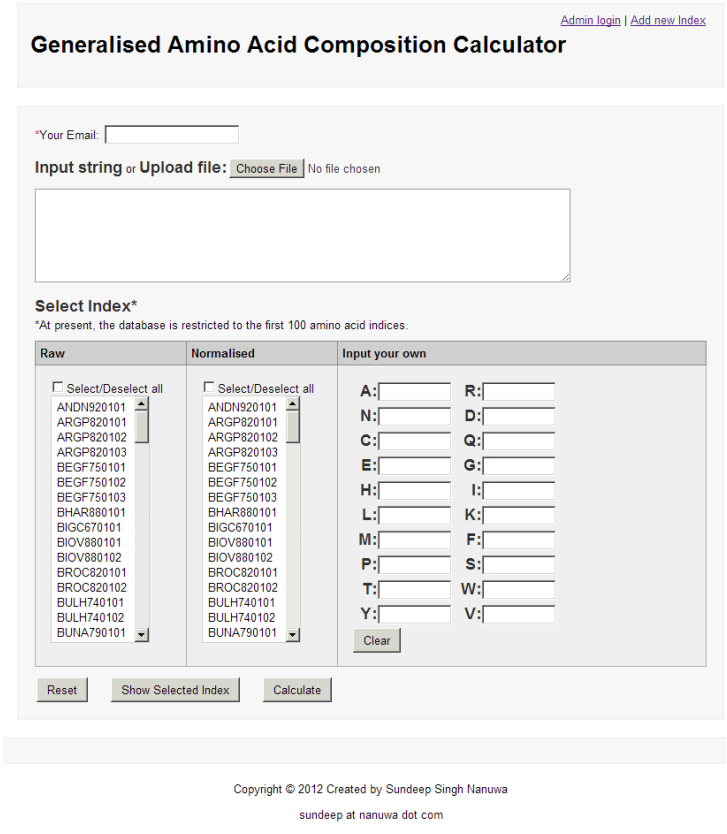
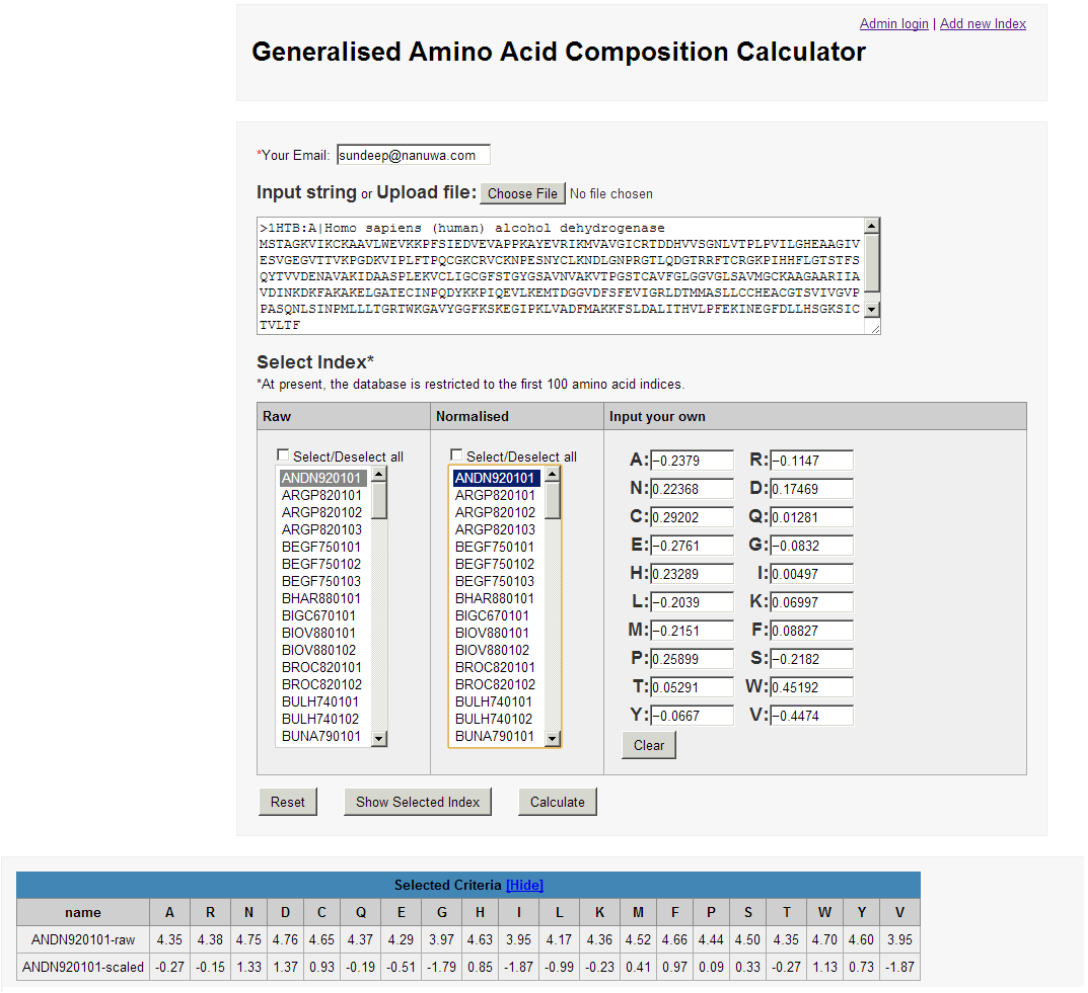


Figure 5-11 Front end of GAAC web server

After a sequence is entered or uploaded, the user then selects which index or indices (available in the database) to convert the sequence into by selecting the various (or all) amino acid indices from the scroll boxes. The user also has the option to select raw and/or normalised amino acid indices values. The user can view the indices values before pressing calculate by pressing show-selected index, which is viewable at the bottom of Figure 5-12. The user is also able to convert protein sequence using user defined index values as show in Figure 5-12.



Selected Criteria [Hide]

name	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
ANDN920101-raw	4.35	4.38	4.75	4.76	4.65	4.37	4.29	3.97	4.63	3.95	4.17	4.36	4.52	4.66	4.44	4.50	4.35	4.70	4.60	3.95
ANDN920101-scaled	-0.27	-0.15	1.33	1.37	0.93	-0.19	-0.51	-1.79	0.85	-1.87	-0.99	-0.23	0.41	0.97	0.09	0.33	-0.27	1.13	0.73	-1.87

Figure 5-12 GAAC webserver populated with initial data

When the user is ready to precede further they must press calculate to generate the data. For each of the selected amino acid indices form the scroll boxes and the user defined set of index values, the converted data is viewable and ready to download as shown at the bottom Figure 5-13.

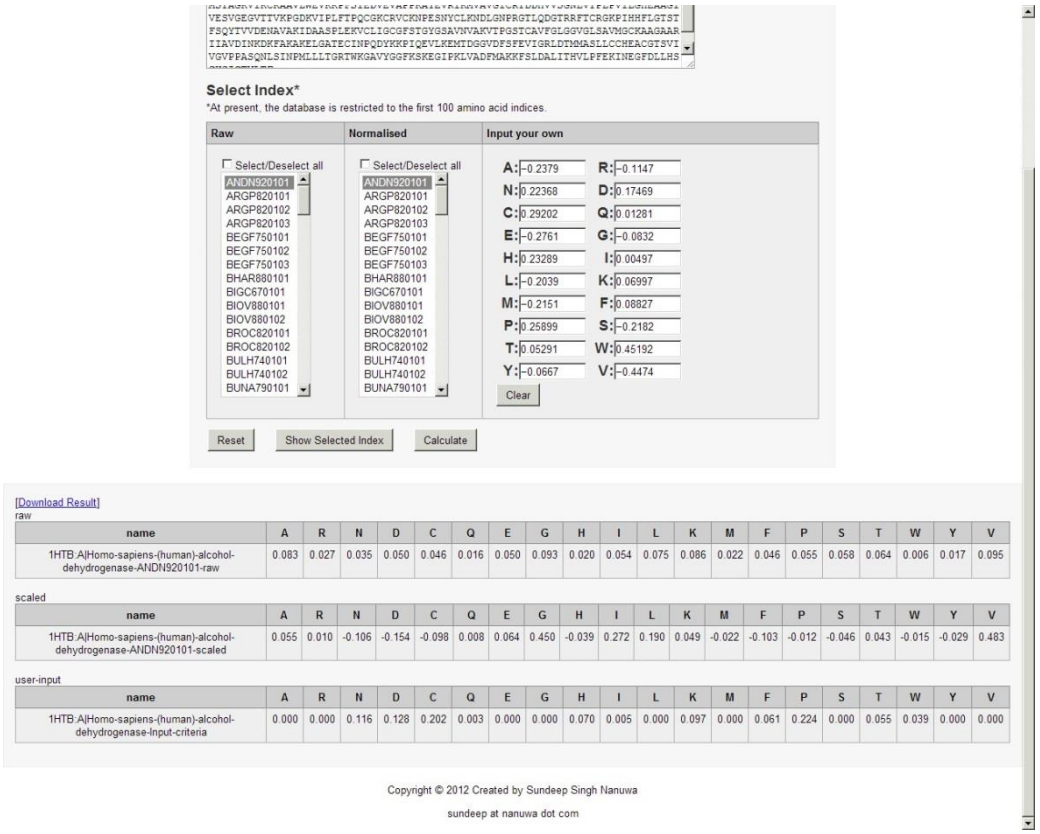


Figure 5-13 Results generated after pressing calculate

The GAAC webserver also allows users to submit new amino acid indices as shown in Figure 5-14. After pressing submit the backend end database will check if it exist in the database already, if so it will inform the user otherwise it will accept it for webserver administration verification.

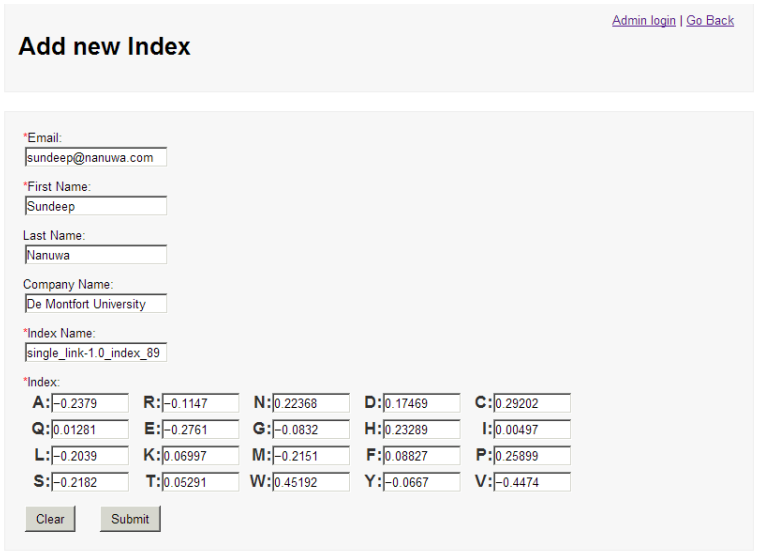


Figure 5-14 GAAC webserver adding new index

5.7 Conclusions

Amino acid indices are a set of features representing many different physicochemical and biochemical properties of amino acids, chapter 5 has presented several novel methods utilising them. The GAAC method presented in this chapter builds upon the identification of one of the strongest composition based feature group predictor in chapter 4 the amino acid composition. The GAAC generalises the amino acid composition and replaces the artificial weight of one with natural amino acid weights by utilising amino acid index values. Results obtained through the GAAC method and amino acid indices has improved the predictive accuracies compared to chapter 4 results – see Table 5-12 to Table 5-14. The six amino acid indices that produced the highest predictive accuracies are AAI # 31 (CHAM830108), 64 (DAYM780101), 414 (AURR980115), 437 (KUMS000101), 466 (FUKS010109) and 568 (Nm). AAI # 568 is one of the newly added amino acid indices that were found through the literature searches carried out. The AAIndex1 is the original source of the majority of this thesis amino acid indices dataset, the databases last update was in March 2008 and the oldest entries are from 2005. This highlights the importance of keeping an amino acid dataset updated as new amino acid indices are regularly published but authors have no central database to deposit the data, thus there is a chance other studies are missing out the potential capabilities of unknown amino acid indices like the identification of AAI # 568. The GAAC method is available at www.generalised-protein-sequence-features.com, which is publically available and allows anyone to generate GAAC sequence driven features and contribute towards developing the amino acid indices database and keeping it up to date.

The amino acid indices dataset contained amino acid indices of varying amino acid properties; however, many of these have been developed for the same amino acid property and these indices have either (1) varying range of descriptor values or very alike range of values this is also true for other different amino acid indices. Regardless of the biochemical property an index is derived for, the dataset contains redundant information that can be summarised. The aim was to reduce this redundancy by clustering similar amino acid indices and then applying PCA to reveal a novel set of index values from each cluster. Many clusters were formed but only the ones that extract a novel index that had a minimum of >0.99% variance in the first principal component was kept for analysis. These novel indices were then used as weights with the GAAC method to derive sequence driven features for the 25PDB and 1189 datasets. The analysis carried out using the new computationally generated amino indices revealed

better predictive accuracies than amino acid indices it replaces. The overall result is that amino acid indices provide, superiority, generality and applicability of which have been tested using protein structural classes with positive outcomes see Table 5-7and Table 5-8 for results.

The results presented in chapter four and five are based on a the full use of the feature space i.e. analysing an amino acid index uses a feature space size of 20 or analysis an dipeptide composition uses a feature space size of 400. This feature space may contain redundant or irrelevant features, which contribute little or no information useful for the prediction analysis in hand and may add noise to the relevant feature space. Identifying and removing these redundant or irrelevant features may (1) improve prediction accuracy (2) reduce the computational analysis time and (3) generalise the feature space. The difference between feature exaction presented throughout chapter 5 and feature selection is that feature extraction creates new features from an original feature space, whereas feature selection returns a subset of the original feature space that is more representative of the original feature space, this area in bioinformatics is feature selection and is explored further in chapter 6.

Chapter 6 - Feature selection

6.1 Introduction

Feature selection is the process of identifying a small but representative subset of the original feature space. The question behind using feature selection is, given a set of features, can a subset selection of the original feature space lead to a better classification performance (Smialowski, Frishman et al. ; Ron and George 1997; Cohen, Tian et al. 2002; Ding, Peng et al. 2003). Using all the features in a dataset often introduces noise and long analysis time, which may reduce the classification accuracies and performance. Feature selection can help identify features that are more representative of the larger feature space, which if removed and analysed, may improve classification performance and uses less computational resources and yield a more compact and representative subset of the original feature space (Cohen, Tian et al. 2002; Saeys, Inza et al. 2007; Hua, Tembe et al. 2009). Chapter 6 is a review of bioinformatics approaches towards feature selection and evaluates two widely methods over the sequence driven features used in chapters 4 and generated in chapter 5.

6.2 Feature selection categories

There are four categories that have been considered in the literature for feature selection (1) filter method category, (2) wrapper method category and (3) embedded method category (Saeys, Inza et al. 2007). Table 6-1 lists the widely reported feature selection methods along with their advantages and disadvantages (Saeys, Inza et al. 2007).

Table 6-1 Feature selection techniques – adopted from (Saeys, Inza et al. 2007)

FS Category	Advantages	Disadvantages	FS Examples	Reference
	Univariate			
Filter	• Fast • Scalable • Independent of the classifier	• Ignores feature dependencies • Ignores interaction with the classifier	F-select	(Chang and Lin 2001)
	Multivariate			
	• Models feature dependencies • Better computational complexity than wrapper methods • Independent of the classifier	• Slower than univariate techniques • Ignores interaction with the classifier • Less scalable than univariate techniques	mRMR	(Peng, Long et al. 2005)
			Fast correlation-based feature selection	(Yu and Liu 2004)
Wrapper	Deterministic			
	• Simple • Models feature dependencies • Less computationally intensive than randomized methods • Interacts with the classifier	• Risk of over fitting • Classifier dependent selection • More prone than randomized algorithms to getting stuck in a local optimum (greedy search)	Sequential forward selection	(Chen 1978)
			Sequential backward elimination	(Chen 1978)
	Randomized			
	• Less prone to local optima • Interacts with the classifier • Models feature dependencies	• Computationally intensive • Classifier dependent selection • Higher risk of over fitting than deterministic algorithms	Genetic algorithms	(Hooker 1995)
Embedded	• Interacts with the classifier • Better computational complexity than wrapper methods • Models feature dependencies	• Classifier dependent selection	Decision trees	(Saeys, Inza et al. 2007)
			Feature selection using the weight vector of SVM	

The filter category does incorporate any classification training (Ni and Liu 2004; Saeys, Inza et al. 2007) but each feature is given a score and depending on the given score criteria features are "filtered" for removal, leaving behind a subset of features that have a high score value ready for analysis. There are two approaches to the filter category are (1) univariate approach, which ignores feature dependencies, and the (2) approach is called multivariate, which considers feature dependencies when scoring.

The wrapper method uses a feature selection method to assess a subset of selected features and then uses a classifier to analysis the selected subset features for a predictive performance. However, the feature selection and classifier are dependent on each other, it may not be the

best methods for the task at hand (Ron and George 1997; Saeys, Inza et al. 2007). The embedded category incorporates a feature selection method inside the classifier. The classifier will continually analyse subset of the features until it gives the best accuracy and penalises the usage of redundant features (Saeys, Inza et al. 2007; Lopes, Martins Jr et al. 2008), but, this approach leads to higher computational analysis time as the large feature space is continually searched for an optimum result.

With consideration of the current feature selection techniques available, the methods employed are f-select and mRMR methods. Both methods are independent of any classifier and are fast, f-select will consider each descriptor value individually with an f-score as it is a univariate based method and mRMR will select top-ranking features based on mutual information between groups of features, as it is multivariate-based method.

6.3 F-select

F-select is a feature selection method that measures the discrimination between two sets of features and is from the univariate filter family methods, which does ignores feature dependencies (Chang and Lin 2001; Peng, Long et al. 2005; Xu, Liu et al. 2008). F-select over dipeptide composition will discriminate each of the 400-descriptor values independently, when inherently dipeptide composition as whole describes 100% of the protein sequence (Li, Lin et al. 2006). However, f-select is useful if the features are not mutually dependent with each other such as amino acid indices where it is represented as a single value through the feature extraction methods using the mean and PCA. The measurement of discrimination is given as a numerical value, which is called an f-score. The larger the f-score value, the more likely the feature is representative towards the target classification variable. The formulas presented are from the supplementary material available at http://www.csie.ntu.edu.tw/~cjlin/.

The f-select algorithm is defined as a training dataset of features $x_i,\ i = 1, \dots, m$; if the number of positive and negative features defined as n_+ and n_-, respectively, then the f-score of the ith feature is defined by in Eq 6-1 as (Chang and Lin 2001):

$$fscore(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(\bar{x}_{k,i}^{(-)} - \bar{x}_i^{(-)})^2}, \quad \textbf{Eq 6-1}$$

Where $\bar{x}_i, \bar{x}_i^{(+)}, \bar{x}_i^{(-)}$ are the average values of the ith feature of the entire dataset, respectively; $\bar{x}_{k,i}^{(+)}$ is the ith feature of the kth positive instance and $\bar{x}_{k,i}^{(-)}$ is the ith feature of the kth negative instance (Lin 2005).

The selection of features obtained through f-select method will be classified using the mknn classifier over the four datasets (25PDB, 1189, Astral25 and Astral40) and evaluated using all three-test procedures (10-fold, leave-one-out and independent-sets).

6.4 Minimum redundancy maximum relevance feature selection

The minimum redundancy maximum relevance feature selection method is a member of the multivariate filter category of feature selection methods and is capable of modelling feature dependencies. mRMR ranks the features according to their maximum relevance to the target classification and the minimum redundancy among the features themselves where the prediction capability has been included by the already selected features (Peng, Long et al. 2005; Li, Lin et al. 2008).

The mRMR algorithm is described below in Eq 6-2 to Eq 6-7. The dataset of all features is defined as *Ω* and the selected subset of the features is defined as *S.* To measure the relevance and redundancy, the mutual information (*MI)* of two features *x* and *y* is defined in Eq 6-2, (MI) is defined to estimate how one feature is related to another (Peng, Long et al. 2005; Li, Lin et al. 2008):

$$MI(x,y) = \sum_{i,j} Prob(x_i, y_j)\, log \frac{Prob(x_i, y_j)}{Prob(x_i)Prob(y_j)} \quad \textbf{Eq 6-2}$$

where $Prob(x_i, y_j)$ is the joint probabilistic distribution of feature x_i and feature y_j; $Prob(x_i)$ and $Prob(y_j)$ are the marginal probabilities of x_i and y_j, respectively (Peng, Long et al. 2005; Li, Lin et al. 2008). Where joint probabilistic distribution refers to, two random features x_i and y_j is defined in the same probability space. The minimum redundancy is defined in Eq 6-3:

$$\min_{S \subseteq \Omega} W_{MI,} W_{MI} = \frac{1}{|S|^2} \sum_{i,j \in S} MI(x_i, x_j) \quad \textbf{Eq 6-3}$$

where $|S|$ is the number of features in the subset of the features *S* and W_{MI} is the minimum redundancy value (Peng, Long et al. 2005; Li, Lin et al. 2008). For the targeted class labelled

$c = \{c_1, c_2, c_3, c_4\}$ the relevance between feature $_i$ and the targeted class label c can be counted by the mutual information $MI(c, x_i)$ between class c and the feature variable x_i (Peng, Long et al. 2005; Li, Lin et al. 2008). The maximum relevance is defined in Eq 6-4:

$$\max_{S \subseteq \Omega} V_{MI,} V_{MI} = \frac{1}{|S|} \sum_{i \in S} MI(c, x_i) \quad \textbf{Eq 6-4}$$

where V_{MI} is the maximum relevance value.

After calculating the mutual information (Eq 6-2), minimum redundancy (Eq 6-3) and maximum relevance values (Eq 6-4), mRMR further optimises Eq 6-3 and Eq 6-4 by (1) selecting a single maximum relevant feature according to Eq 6-4, i.e. select feature i such that $MI(c, x_i)$is higher than other features. This further optimisation is called maximum relevance optimisation and is defined in Eq 6-5 (Peng, Long et al. 2005; Li, Lin et al. 2008):

$$\max_{i \in \Omega} MI(x_i, c) \quad \textbf{Eq 6-5}$$

where Ω is the whole feature set (Peng, Long et al. 2005; Li, Lin et al. 2008). The remaining features are selected by adding an additional feature i to *S* to satisfy either of the two-mRMR equations as defined in Eq 6-6 and Eq 6-7 (Peng, Long et al. 2005; Li, Lin et al. 2008):

$$\max_{i \in \Omega_S} [MI(c, x_i) - \frac{1}{|S|} \sum_{j \in S} MI(x_i, x_j)] \quad \textbf{Eq 6-6}$$

$$\max_{i \in \Omega_S} [MI(c, x_i) / \frac{1}{|S|} \sum_{j \in S} MI(x_i, x_j)] \quad \textbf{Eq 6-7}$$

where $\Omega_s = \Omega - S$, represents the features still to be selected. Eq 6-6 and Eq 6-7 are called the mutual information difference (MID) selection criteria and mutual information quotient (MIQ) selection criteria, respectively (Peng, Long et al. 2005; Li, Lin et al. 2008).

The selection of features obtained in this chapter through the mRMR feature selection method will be classified using the mknn classifier over the four datasets (25PDB, 1189, Astral25 and Astral40) and evaluated using three test procedures (10-fold, leave-one-out and independent-sets). F-select and mRMR implementations were provided as MATLAB software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/ and http://penglab.janelia.org/proj/mRMR/, respectively.

6.5 Results from feature selection methods

The following three sequence driven features were processed by f-select and mRMR feature selection method to obtain a smaller set of input feature vector.

(1) Traditional sequence driven features for each dataset presented in chapter 4

(2) 611 extracted features using the SRM utilising the mean presented in chapter 5

(3) 611 extracted features using SRM utilising PCA presented in chapter 5

The goal was to reduce the feature size and potentially improve the prediction accuracy when compared to using all features together or feature groups. The feature selection methodology was performed in two steps: (1) feature selection methods was applied to select a subset of the most relevant features and (2) the selected features were classified to determine if the predictive accuracy is better than using what is presented in chapter 4 and 5 results. In the interest of keeping the results concise, only the top 10 selected features are shown and analysed.

6.5.1 Feature selection results over the traditional sequence driven features presented in chapter 4

- The traditional sequence-driven feature set presented in chapter 4 includes 10 feature groups, with 1497 descriptor values and 53 subsets of the features, f-select and mRMR methods were applied to the whole feature set.
- Table 6-2 contains the top 10 selected features over each dataset and highlights the comparisons between the two feature selection methods and overall, both feature selection methods are quite similar with in terms of selected features and rank order. The interesting outcome of the feature selection result is the majority of selected features centre around feature index 6.1.6 – secondary structure that forms the composition feature group (feature index 6.1). Both methods selected feature index 1156 as the most discriminative feature, index 1156 is the 1st descriptor value of sub feature group 6.1.6 which is the secondary structure composition which contains three descriptors helix, strand and coil, which are 1156, 1157 and 1158 respectively. Feature group 6.1.6 appears consistently in top 10 results in chapter 4, as 9 out of 12 sets of analyses ranked feature it in the top 10 as shown in Table 4-13. Moreover, the feature's biological references are toward the composition of secondary structure elements, which links closely with chapter 4 conclusions that composition based

features are better suited for the prediction of protein structural classes. The commonly selected feature across each dataset is shown in

- Table 6-2 are (refer to appendix I for the feature index ranges):-

- 1189 dataset, first and second selected features are 1156 and 1157
- 25PDB dataset, first selected feature is 1156
- Astral25 dataset, first and second selected features are 1156 and 1158
- Astral40 dataset, first and second selected features are 1156 and 1158

Table 6-2 Top 10 selected features using f-select and mRMR over the 1497 sequence-driven features (refer to appendix I for feature names and ranges)

Dataset	25PDB		1189		Astral25		Astral40	
Rank	f-select	mRMR	f-select	mRMR	f-select	mRMR	f-select	mRMR
1	**1156**	**1156**	**1156**	**1156**	**1156**	**1156**	**1156**	**1156**
2	902	1157	1157	1157	1158	1158	1158	1158
3	1157	1141	1158	1179	902	1141	662	1141
4	1141	1160	1179	1158	1141	1160	902	1160
5	1160	1321	1141	1149	1160	1149	422	1157
6	1149	1320	1160	1159	1157	1157	1157	1149
7	1179	1158	18	1141	1149	1179	1149	1179
8	1158	1153	1149	1160	662	1142	1141	1142
9	424	1179	902	1153	422	10	1160	1148
10	663	1149	662	1251	6	1144	6	1159

Results presented in Table 6-4 to Table 6-9 are the predictive accuracies obtained using f-select and mRMR, incrementally adding the next selected features to analyse. Each table contains the predictive accuracies, where the column named “number of features” indicates how many of the selected features were used for analysis. Within the tables, figures that are highlighted bold are highest result obtained for each analysis. Two results obtained from f-select and six obtained from mRMR generated higher accuracies than what was presented in Table 4-13, results are shown in Table 6-3 with comparison to chapter 4, the range of increase in accuracy is between 0.04% - 1.43%. The column “New Feature Size” refers to the number of features used to obtain the highest accuracy and the selection of features corresponds to Table 6-2, e.g. the first row in Table 6-3 the “New Feature Size” is 5, in Table 6-2 the first 5 select features in column “25PDB / f-select” are 1156, 902, 1157, 1141 & 1160.

Table 6-3 Highest results obtained from each feature selection

Feature Selection Method	Dataset / Test procedure	New Feature Size	Highest Accuracy	Previous highest accuracy	Old Feature Size	Feature Index (see appendix I)	Increase
F-select	25PDB / 10-fold	5	41.73%	41.69%	20	1	0.04%
F-select	25PDB / LOO	6	41.97%	41.91%	20	1	0.06%
mRMR	25PDB / 10 fold	6	43.12%	41.69%	20	1	1.43%
mRMR	1189 / 10-fold	5	41.82%	41.42%	400	2	0.4%
mRMR	Astral40 / 10-fold	10	40.57%	40.33%	21	6.1	0.24%
mRMR	25PDB / LOO	6	42.87%	41.91%	20	1	0.96%
mRMR	1189 / LOO	5	42.58%	41.84%	400	2	0.64%
mRMR	Astral40 / Independent-sets	9	42.56%	41.38%	20	1	1.18%

Table 6-4 Prediction accuracies obtained using F-select with 10-fold test procedure and MKNN classifier

Number of features	25PDB	1189	Astral25	Astral40
1	39.33%	36.41%	36.18%	35.72%
2	41.17%	39.26%	36.27%	36.21%
3	39.46%	39.76%	36.53%	36.23%
4	40.28%	38.63%	38.93%	36.26%
5	**41.73%**	38.71%	39.24%	36.66%
6	41.19%	39.82%	38.41%	37.07%
7	41.55%	**40.02%**	38.63%	38.81%
8	40.63%	38.63%	38.76%	38.82%
9	40.63%	38.63%	38.86%	38.28%
10	40.69%	38.72%	**39.73%**	**39.61%**

Table 6-5 Prediction accuracies obtained using F-select with leave-one-out test procedure and MKNN classifier

Number of features	25PDB	1189	Astral25	Astral40
1	40.05%	36.59%	36.66%	35.46%
2	41.55%	38.34%	36.18%	36.32%
3	40.17%	38.89%	36.62%	36.47%
4	41.43%	38.80%	39.49%	36.55%
5	41.79%	38.71%	39.70%	36.76%
6	**41.97%**	39.36%	38.81%	37.51%
7	41.55%	**40.37%**	39.53%	39.20%
8	41.07%	39.54%	39.57%	39.57%
9	41.13%	39.54%	39.57%	39.16%
10	41.19%	39.54%	**40.94%**	**40.61%**

Table 6-6 Prediction accuracies obtained using F-select with independent-sets test procedure and MKNN classifier

Number of features	25PDB	1189	Astral25	Astral40
1	45.74%	41.11%	36.93%	36.96%
2	45.56%	43.32%	37.25%	36.93%
3	46.58%	44.15%	37.68%	37.31%
4	47.72%	45.35%	38.99%	37.54%
5	47.24%	45.99%	39.48%	40.34%
6	47.78%	45.07%	39.69%	40.68%
7	49.22%	**48.02%**	40.36%	**41.11%**
8	49.22%	47.47%	40.34%	40.95%
9	49.28%	47.47%	40.45%	40.96%
10	**50.42%**	47.47%	**40.79%**	41.04%

Table 6-7 Prediction accuracies obtained using mRMR with 10-fold test procedure and MKNN classifier

Number of features	25PDB	1189	Astral25	Astral40
1	39.33%	36.41%	36.18%	35.72%
2	38.82%	39.26%	36.27%	36.21%
3	40.29%	38.25%	39.10%	38.00%
4	41.67%	38.63%	39.03%	37.64%
5	41.24%	**41.82%**	39.45%	37.85%
6	**43.12%**	40.25%	38.61%	38.20%
7	41.62%	39.34%	40.36%	39.17%
8	41.42%	38.95%	40.27%	39.63%
9	42.00%	39.86%	40.65%	39.98%
10	42.36%	41.65%	**40.86%**	**40.57%**

Table 6-8 Prediction accuracies obtained using mRMR with leave-one-out test procedure and MKNN classifier

Number of features	25PDB	1189	Astral25	Astral40
1	40.05%	36.59%	36.66%	35.46%
2	38.73%	38.34%	36.18%	36.32%
3	41.19%	38.16%	39.40%	38.37%
4	41.97%	38.80%	39.68%	38.55%
5	42.33%	**42.58%**	40.29%	38.57%
6	**42.87%**	40.55%	39.43%	39.12%
7	41.79%	39.72%	40.10%	39.50%
8	41.97%	38.80%	40.37%	40.13%
9	42.09%	40.37%	40.97%	40.22%
10	42.81%	42.30%	**41.01%**	**41.39%**

Table 6-9 Prediction accuracies obtained using mRMR with independent-sets test procedure and MKNN classifier

Number of features	25PDB	1189	Astral25	Astral40
1	45.74%	41.11%	36.96%	36.93%
2	46.58%	43.32%	37.23%	37.25%
3	47.48%	44.70%	39.28%	39.16%
4	48.20%	45.35%	39.39%	39.14%
5	49.22%	46.18%	38.99%	39.75%
6	51.26%	48.66%	39.21%	40.22%
7	51.02%	47.83%	40.40%	40.66%
8	53.36%	48.39%	40.34%	41.30%
9	55.52%	48.94%	40.09%	**42.46%**
10	**56.06%**	**50.97%**	**41.07%**	42.38%

6.5.2 Feature selection results based on the sequence driven features presented in chapter 5 – method 3

This sequence driven feature dataset contains 611 features – each one represent an amino acid index. Table 6-10 contains the top 10 selected features over each dataset and highlights the comparisons between the two feature selection methods, see Appendix II for amino acid index numbers. The frequently highest selected amino acid index is AAI # 568, which appears in each of the top 10-selected feature across each dataset and feature selection method, also ranks the highest in terms of feature selection scores 4 out of 8 times. The selection of features remains consistent across all datasets and feature selection methods; however, the rank order of the top 10 selected features differs across each dataset and feature selection method, which suggests that the differences in protein samples between datasets influences the rank order. Classification accuracy obtained using the first selected feature corresponds to the individual accuracy obtained; example, using the 25PDB dataset with AAI 568 # achieves an accuracy of 42.26% (which is the same obtained in Table 5-22 in chapter 5). Classifying using the top two selected features AAI # 568 and 415 achieves 38.56% and as the number of features increases the accuracy drops further top 10 selected features obtain an accuracy of 37.75%.

Table 6-10 Comparison of Selected features between f-select and mRMR for method 3 (mean) (refer to appendix II for AAI names)

Dataset	F-select				mRMR			
Rank	25PDB	1189	Astral25	Astral40	25PDB	1189	Astral25	Astral40
1	568	412	97	568	568	412	97	568
2	415	222	568	222	415	222	568	222
3	599	261	230	230	414	568	305	230
4	140	568	265	97	599	408	230	97
5	91	408	412	140	140	414	265	265
6	414	414	415	160	91	230	412	412
7	326	230	305	119	326	410	415	140
8	162	410	162	100	162	261	162	160
9	254	19	222	265	254	19	222	119
10	100	100	254	412	100	100	254	100

6.5.3 Feature selection results based on the sequence driven features presented in chapter 5– method 4

Much like the previous section, the feature selection results are based on the feature extraction methods utilising the amino acid indices, which is represented by PCA as presented in chapter 5. This sequence driven feature dataset contains 611 features – each one representing an amino acid index. Similarly, Table 6-11 contains the top 10 selected features over each dataset and highlights the comparison between the two feature selection methods,

see appendix II for amino acid index numbers. The frequently highest selected amino acid index is AAI 166, which appears in each of the top 10-selected feature across each dataset and feature selection method; it also ranks the highest in terms of feature selection scores 6 out of 8 times. Again, similar to the previous section the selection of features remains consistent across all datasets and feature selection methods; however, the rank order of the top 10 selected features differs across each dataset and feature selection method. The classification accuracies obtained using the top 10 selected features did not reach levels of accuracy obtained in Table 5-24. Classification accuracy obtained using the first selected feature corresponds to the individual accuracy obtained; example using the 1189 dataset with AAI # 45 achieves an accuracy of 42.79% (which is the same obtained in Table 5-22 in chapter 5). Classifying using the top two selected features AAI # 45 and 166 achieves 40.89% and as the number of features increases the accuracy drops further top 10 selected features obtain an accuracy of 40.30%.

The classification accuracies obtained using the top 10 selected features did not reach anything higher than what is shown in Table 5-23 and Table 5-24, this was the reason results were not presented in thesis.

Table 6-11 Comparison of Selected features between f-select and mRMR for method 4 (PCA) (refer to appendix II for AAI names)

Dataset	F-select				mRMR			
Rank	**25PDB**	**1189**	**Astral25**	**Astral40**	**25PDB**	**1189**	**Astral25**	**Astral40**
1	**166**	45	**166**	**166**	**166**	**166**	**166**	81
2	553	**166**	121	581	78	553	121	**166**
3	107	396	287	128	553	396	287	128
4	104	167	288	327	107	167	167	578
5	40	51	7	521	104	51	288	327
6	403	341	253	344	40	15	7	521
7	396	553	276	578	403	37	253	344
8	295	333	167	267	396	341	276	267
9	255	15	600	487	295	45	600	487
10	78	37	295	170	255	333	295	170

6.6 Conclusions

Feature selection aims to improve predictive accuracy by selecting a subset of features that is at least equal or more representative than the original feature space, thus, finding a smaller feature space allows a more refined feature set that is less computationally resourceful to analyse. In this chapter, two-feature selection methods have been explored, f-select and mRMR, in which overall both methods resulted in a similar set of results in terms of selected subsets of features. The investigation shows that in the context of protein structural classes

feature selection did not add any significant increase in classification accuracies. However, it did highlight features 1156, 1157 and 1156 as the most representative subsets, which forms the secondary structure composition of a protein sequence – which is biologically related towards the structural classes of protein capturing helix, strand coil structure compositions.

Feature extraction methods 3 and 4, each resulted in a different sets of representative amino acid indices; method three most representative is AAI # 166 and method 4 is AAI # 568. AAI # 166 detects antiparallel and parallel beta-strands that differ in amino acid residue and AAI # 586 detects average medium contacts between amino acid residues.

The results have shown that feature selection is a viable method for selecting a subset of refined features whilst keeping intact as much information as possible but the increase accuracy is at most by 1.43%. The secondary structure sub feature group is the frequently selected feature, which is due to its biological importance towards protein structural classes. However, the marginal differences between feature selection results and highest results obtained without feature selection shows that the features removed contains information required to achieve classification accuracy on par without feature selection. The positive angle of the results obtained is the actual selections of features were more of an interest because it correlates with the top 10 ranked features. This can be considered as a form feature selection as we can eliminate noisy feature groups or amino acid indices based on classification accuracies of the top 10 and disregarding the rest. Overall, feature selection based on the feature reduction methods did not return any significant results in terms of predictive accuracy. However, the selection order of features represented the highest predicted features without feature selection.

Chapter 7 – Discussion, conclusion and future work

7.1 Introduction

Chapter 7 summarises the work described in the previous chapters, it looks at the results, the findings presented and draw conclusions on the relationships between them towards the prediction of structural classes of proteins, and if it has added any value towards the field. Towards the end of the chapter, the key aspects of the thesis are summarised and a section on future work is presented.

7.2 Critical evaluation of traditional sequence-driven features

Chapter 4 presented the analysis of the largest set of traditional sequence-driven features, composed of ten feature groups, 53 sub feature groups with 1497 descriptor values. The analysis of the traditional sequence-driven features had three purposes (1) carry out the largest analysis of sequence-driven features for the prediction of structural classes of protein, (2) to identify which sequence-driven features are better suited for the prediction of structural classes of protein and (3) use the results as a benchmark for chapter 5 work. The results obtained from each sequence driven feature group are further discussed in the following sections.

7.2.1 Composition based sequence-driven-feature groups

There are three feature groups based on the composition of protein sequences, the composition is the fraction of each amino acid or a certain property that appears in the protein amino acid sequence. Traditional composition based feature groups are feature index 1 - amino acid composition, feature index 2 - dipeptide composition and feature index 6.1 - composition. It has been shown in the literature that amino acid composition is a strong predictor of protein structural classes (Eisenhaber, Frömmel et al. 1996; Roy, Martinez et al. 2009; Ahmadi Adl, Nowzari-Dalini et al. 2012). The result presented in chapter 4 confirms that and that the other composition based feature groups are also strong predictors. The amino acid composition feature group was the highest ranked feature 4 out of 12 analyses, which

generated the highest predictive accuracies. However, feature index 2 - dipeptide composition resulted in highest number of ranked results 5 out of 12 times and feature index 6.1 - composition feature group resulted in highest ranked feature 3 out of 12. From all the main sets of analyses (see Table 4-13), each set's highest predicted feature is from a composition based feature group. Furthermore, to support the strength of the composition based feature groups is that the 2nd highest ranked features is usually another composition based feature group, example the first ranked feature of 25PDB / 10-fold analysis is feature index 1 - amino acid composition the 2nd ranked feature is the feature index 6.1 - composition.

7.2.2 Autocorrelation feature groups

The autocorrelation sequence-driven feature groups define the correlation between protein amino acid sequences in relation to a specific structural or physicochemical property (Broto, Moreau et al. 1984). Autocorrelation sub features are amino acid indices, which are used to derive the descriptor values, however the selection of amino acid indices were pre-defined by the authors of the feature group (Li, Lin et al. 2006) and the drawback is that it only uses a eight indices out of the 611 available amino acid indices. The main observation found is with the hydrophobicity scale, which was developed in the context of protein structural classes studies. It produces the majority of the highest predictive accuracies for each of the autocorrelation feature groups. The biological significance of the hydrophobicity scale is the property of a protein being water-repellent, tending to repel and not absorb water (Tanford 1962; Bigelow 1967; Charton and Charton 1982; Horne 1988; Kumarevel, Gromiha et al. 2000). The hydrophobicity scale takes into account the environment of each amino acid residue when estimating its hydrophobicity value, which is dependent on the structural class present in the protein. It is believed that the datasets contain proteins samples that represent structural classes, is the reason why the hydrophobicity sub feature consistently appears in the top 10 sets of summarised results in Table 4-13. Using independent-sets test procedure, where training datasets are Astral25 and Astral40, the majority of the top 10 ranked features are from the autocorrelation-based feature groups, which seems like the autocorrelation feature groups and its sub features are more suited when larger datasets are used as training. Whereas 10-fold and LOO top 10 ranked features are mainly from the composition-based feature groups.

7.2.3 Composition, transition and distribution feature groups

Feature index 6.1, 7.1 and 8.1 are the composition, transition and distribution feature groups, respectively. All three-feature groups have the same set of seven physicochemical properties used to compute its descriptor values. Results show that the transition and distribution of amino acids along a protein amino acid sequence does not define protein structural classes very well compared to composition based feature group. None of the feature group or sub-feature indexes between 7.1 and 8.1.7 appeared in the top 10 ranked features as the predictive accuracies for the feature groups is consistently lower than feature groups 6.1. The strongest feature group and sub feature are feature index 6.1 and 6.1.6, composition feature group and composition - secondary structure sub feature, respectively, feature index 6.1.6 is a sub feature of feature index 6.1-composition feature group. The biological link between sub feature 6.1.6 – secondary structure and prediction of protein structural classes is the calculation of secondary structure elements helix, strand and coil (Ahmadi Adl, Nowzari-Dalini et al. 2012). Sub feature group 6.1 – composition is the global percentage of each encoded class in the sequence (helix, strand and coil) (Lin and Pan 2001). These encoded classes are the fundamental aspects of secondary structure elements, which, what protein structural classes' prediction aims to find out - the majority secondary structural element. The composition sub-feature group uses the same equation as amino acid composition, instead of calculating composition of amino acids; it calculates composition for each encoded class in the sequence. What has been found from the research is protein structural class prediction is more about what the sequence is composed of rather than how often the sequence goes into different states (transition) and how the sequence is distributed at different positions (distribution and sequence order).

7.2.4 Pseudo amino acid composition

Pseudo amino acid composition incorporates amino acid composition and sequence-order information. The sequence order descriptor values are derived from three amino acid indices the hydrophobicity, hydrophilicity, and side-chain mass amino acid indices. Again, pseudo amino acid composition uses a limited number of amino acid indices, compared to the size of the amino acid indices database. Results show that the majority of the predictive power comes from the first sub-feature amino acid composition and that sequence-order effect does not add much more to the prediction accuracy. Pseudo amino acid composition analysis was split into three different groups (1) the whole Pseudo amino acid composition feature index 10

(50 descriptors) (2) AAC feature index 10.1 (20 descriptors) and (3) lambda feature index 10.2 (30 descriptors). For each testing dataset using independent –set test procedure, sub feature index 10.2 results are higher than feature group index 10; this shows that the amino acid composition is what drove the prediction accuracy. With the exception where evaluation methods are 10-fold and leave-one-out and testing datasets are Astral25 and Astral40 (sets 7, 8, 10 and 11 see Table 4-12) feature index 10 accuracies are higher than feature index 10.1 accuracies, which indicates that feature 10.1 sequence-order effects provided marginal support to the feature index 10.2.

7.3 Critical evaluation of amino acid indices based sequence-driven-features

Following from Chapter 4 work the evaluation of the traditional sequence driven features, which has shown that composition based feature groups are the most representative of protein structural classes, where four of the traditional sequence driven feature groups utilised amino acid indices to derive the descriptor values, feature groups 3, 4, 5 and 10, normalized moreau-borto autocorrelation, moran autocorrelation, geary autocorrelation and pseudo amino acid composition, respectively. The selection of amino acid indices used within these feature groups were limited compared to the number amino acid indices that are available for analysis. This thesis utilised the largest set of amino acid indices, best to our knowledge, in a number of different novel ways, the development of these novel methods are discussed in the following sections.

7.3.1 Updated amino acid indices dataset

The original amino acid database named AAindex1 developed by Kawashima, S et al (2000) contained 544 entries as of March 31, 2008 (Kawashima and Kanehisa 2000). The AAindex1 database contains entries that have missing index values and redundant entries. The study started by ensuring the thesis amino acid indices dataset was free of such entries; hence, 16 of them were removed. It then moved on to collecting as many published amino acid indices that have not been previously deposited into the AAindex1 database and 83 additional amino acid indices were added to the amino acid indices dataset through literature searches. The thesis has developed the largest amino acid indices' dataset to date, consisting of 611 amino acid indices and should be considered as the benchmarked dataset for any amino acid indices related study.

7.3.2 Generalised amino acid composition

The link between traditional sequence driven features and generalised amino acid composition (GAAC) method is with the amino acid composition feature group. It is one of the strongest traditional feature groups, thus, the focus turned towards how to utilise it further. A modification to the amino acid composition formula as shown in Eq 4-1 in chapter 4, takes into account the many hundreds of natural amino acids weights that are available in the amino acid indices datasets. The modification replaces traditional amino acid composition normalised weight of one for each amino acid type with an amino acid index. This allows a flexible approach of generalising amino acid composition by utilising the many hundreds of natural amino acids weights available.

The GAAC method improved prediction accuracy results over the initial set of benched mark results (chapter 4) using traditional sequence driven features across all datasets, comparison of results are shown in Table 5-12 to Table 5-20, the increases in results are between 4.15%-9.82% over all analyses. It has shown that replacing the normalised weight in traditional amino acid composition feature group with amino acid indices are more representative of protein structural classes than traditional feature groups analysed because it utilises natural amino acid weights. The largest increases are with using 25PDB and 1189 datasets where the test procedure is independent-sets. Compared to amino acid compsotion feature set, GAAC increase was higher by 6.47% and 10.24%. The reliability of the results is in the selection of amino acid indices that have significant biological references towards prediction of protein structural classes. Amino acid indices have been identified that are more representative towards the four main proteins structural classes, which are discussed in the next section.

7.3.3 Identification of a candidate set of amino acid indices

Utilising the largest collection of natural indices a sub-set of them were found to be highly related to the prediction of structural class of proteins. This section highlights several potential amino acid indices candidates that are useful for predicting the structural class of a protein, these indices are shown in Table 5-11 and the relevance towards protein structural classes discussed here.

Amino acid index 414 (AURR980115) - Normalised positional residue frequency at helix termini C1 (Aurora-Rose, 1998). Helix capping are specific patterns of hydrogen bonding found at the ends of proteins, when certain amino acid residues are exposed to solvents, proteins will

compensate by helix capping to protect the protein. The index values are based on the normalised positional frequencies of a given residues at the helix terminia position C1. The authors examined the impact of helix capping with the conformation of secondary structure (Aurora and Rose 1998). Capping imposes restrictions on the number of conformation helices; this reduces the search spaces and improves the rate of correct conformation.

Amino acid index 31 (CHAM830108) - A parameter of charge transfer donor capability (Charton and Charton 1983). This index calculates the probability that each amino acids type have propensities in different regions of secondary structures. This index is based on the Chou-Fasman method, which is one of the earliest methods used for protein secondary structure prediction. The Chou-Fasman method is a set of probability parameters (also known as values of side chain parameters) for the appearance of each type of amino acid in each secondary structure type. The index values are either +1 or 0 (non-normalised). Amino acids Ala, Asp, Glu, Gly, IIe, Leu, Pro, Ser, Thr, Val are set to zero, which means they are not expressed. Amino acid Arg, Asn, Cys, Gln, His, Lys, Met, Phe, Trp, Tyr are set to one, which means they are expressed. Normalised values are between -1 and +1 which results are obtained from. The amino acid indices look for specific amino acid expressions, which are more prevalent in the datasets, used in the analysis.

Amino acid index 437 (KUMS000101) Distribution of amino acid residues in the 18 non-redundant families of thermophilic proteins (Kumar et al., 2000) (Kumar, Tsai et al. 2000). Index 437 describes the percentage distribution of amino acids in thermophilic proteins. It has been suggested thermophilic proteins have higher helical content and thermostability has also been attributed to enhanced secondary structure propensity (Querol, PerezPons et al. 1996) hence why it is possible this index picks up protein structural class prediction. This index has a high correct rate for α/β proteins. The astral40 datasets contains the largest number of α/β proteins compared to the other datasets, when analysis using LOO and independent test procedures the predictive accuracies are 79.37% and 75.32%, respectively - which is amongst the highest accuracy for α/β proteins.

Amino acid index 466 (FUKS010109) Entire chain composition of amino acids in intracellular proteins of thermophiles (Fukuchi and Nishikawa 2001). Indices are derived for the protein group intracellular proteins of thermophiles, the values represents the average compositions of entire amino acid chains.

Amino acid index 568 – average medium contacts (N_m) (Fernandez, Caballero et al. 2007). The average short, medium and long range contains for the residues in each of the four structural classes' shows that the short range contains are similar in all classes and the role of medium and long-range contacts are distinct in each class. The average medium-range contacts are higher for all alpha class proteins than all beta proteins, indicating the influence of medium range contains in the formation of a-helices. Range contacts are based on the amino acid residues, which are in contact with each other. For a given residue, the composition of surrounding residues is analysed in terms of their location at the sequence level. Residues that are within a distance of two residues from the central residues are considered contribute to short-range interaction, those within a distance of ±3 or ±4residues to medium range and those more than four residues away to long-range interaction (Gromiha and Selvaraj 2004).

7.3.4 Generalised amino acid composition webserver

As discussed GAAC is a better feature group for the prediction of protein structural classes, this led to the development of a webserver that enables the calculation of GAAC based sequence driven feature groups which is available online at http://www.generalised-protein-sequence-features.com/. As well as calculating GAAC, the webserver is intended to be a place to find and deposit new amino acid indices from any user.

7.3.5 Amino Acid Indices based sequence-driven-feature extraction methods

The results of the sequence-driven-feature extraction methods 3 and 4 are presented in chapter 4. The two feature extraction methods did not achieve anything better than the GAAC method, in the context prediction of protein structural classes no additional benefit was generated. This has led to the reason that too much information is lost by extracting features using the mean and PCA. However, there is still the potential that these two feature extraction methods can be useful in other proteomic studies and/or the wider bioinformatics field. Between the two-feature extraction methods, the utilisation of the mean (method 3) is a better method than PCA (method 4), as overall the results were higher. In additional to new feature extraction methods, these methods are explored to identify possible amino acid indices candidates for the prediction of protein structural class of proteins. The selections of top 10 amino acid indices are very different from the GAAC method and the highest predicted amino acid index is 414 at 59.36%.

7.3.6 Hybrid sequence driven feature extraction

The hybrid sequence driven feature extraction used hierarchal clustering to cluster similar amino acid indices together, the threshold value where to cut the hierarchal cluster tree was obtained through some trial and error of different cut off points. Two optimum cut off points 1.0 and 0.65 were found. A higher cut-off point resulted in all indices clustered into a single cluster and a lower cut-off point resulted in each index clustered on its own. Six sets of different cluster methods produced six sets of cluster amino acid indices, the arrangement of clustered indices are available at http://cisaps.com/indices/default/generated_indices. The tables are too big to include in the thesis. However, the result of clustering amino acid indices have reduced the thesis amino acid indices datasets size from 611 to the following:-

- 107 amino acid indices using SINGLE Linkage and Minimum Cluster Distance = 1.0
- 134 amino acid indices using SINGLE Linkage and Minimum Cluster Distance = 0.65
- 181 amino acid indices using COMPLETE Linkage and Minimum Cluster Distance = 1
- 216 amino acid indices using COMPLETE Linkage and Minimum Cluster Distance = 0.65
- 155 amino acid indices using AVERAGE Linkage and Minimum Cluster Distance = 1
- 155 amino acid indices using AVERAGE Linkage and Minimum Cluster Distance = 0.4

Majority of the clustered amino acid indices are following the high cross correlation values presented in the AAindex 1 database (Kawashima and Kanehisa 2000). Looking at the reduced amino acid indices dataset the smallest dataset contains 107 clusters and the largest contains 216 clusters. This shows that there was high rate of redundant data present in the original amino acid indices dataset. Each cluster had PCA applied to it and the first principal component extracted to revel the summarised amino acid index, Table 5-6 shows the number of computationally generated indices with ≥ 0.99 variance. The threshold of 0.99 was chosen to capture any first principal components that fell just below the 1.0 variance threshold. Majority of the computationally generated indices that have a variance of >=0.99 are in fact 1.0, only a handful computationally generated indices are less than 1.0 and >=0.99. When the variance is 1.0 it means that, the computationally generated index represents 100% of original clustered data.

Each of the computationally generated indices were analysed on both 25PDB and 1189 datasets using independent-set test procedure, where training datasets are Astral25 and Astral40 respectively as it produces the most robust set of results. In each analysis, the

predictive accuracy of the computationally derived indices was higher than the original amino acid indices it replaces. Consolidating highly similar amino acid indices had removed noise that was present in the amino acid indices dataset and further refined it thus reducing the search space for finding suitable amino acid indices.

The hybrid method produced the highest accuracy for 25PDB testing dataset where the training dataset was Astral25, results obtained were higher than what has been achieved using the GAAC method. The best computationally derived amino acid indices that produced the highest predictive accuracy came from three different hierarchical clustering methods each with the same predictive accuracy of 75.52%. Refer to appendix IV, V and VII, respectively.

- single linkage cut off point 1.0 - generated index 48,
- single linkage cut off point 0.65 - generated index 105
- complete linkage cut off point 0.65 – generated index 176

Each of the generated indices clustered the amino acid index feature index 409 and 414. Refer to appendix II for amino acid indices information. Feature index 414 is one the amino acid indices identified as suitable candidate (presented in section 7.3.3) that is representative towards predicting of protein structural classes. Index 409 has biological references towards secondary structure element helix. The variance value of each of the cluster is 1.0, which means the computationally derived index represents 100% of the original data.

7.4 Feature selection

Results presented in chapter 6 are the outcome of feature selection analysis; the results highlighted the most descriptive feature from the traditional sequence feature reset being index 6.1.6, which is the composition of alpha helix secondary structure elements, belonging to the secondary structure sub feature. This has a high biological reference towards protein structural class prediction. The result of the classification analysis based on, up to the top 10 selected features did not generate a significant increase in predictive accuracy higher than what has already been achieved in Table 4-13 of chapter 4 results and feature extraction methods 3 and 4, Table 5-23 and Table 5-24 respectively. Both feature selection methods have their advantages and disadvantage; however, it was shown that mRMR is the better of the two as it considers mutual information between features and produced six results higher than Table 4-13 from chapter 4 studies as presented in Table 6-3.

7.5 Test procedures

Test procedures are the classification evaluation methods involved in partitioning the dataset, the three widely used test procedures are n-fold, leave-one-out and independent-sets. All three-test procedure are adopted in the analyses this thesis undertook as most often studies only consider one of the test procedures (Kurgan and Homaeian 2006). Each test procedure produced different sets of ranked features and predictive accuracies; this shows that the choice of test procedures (data partitioning) method affects results and that a consensus from all three methods should be drawn upon and not from just one single method. The 10-fold test procedure is the least computationally intensive but outputs marginally lower accuracies compared to leave-one-out. The leave-one-out is the most computationally demanding but with slightly higher accuracies compared to 10-fold. Independent-sets, in which testing datasets are 25PDB and 1189 and training datasets are Astral25 and Astral40, respectively, produced the highest sets of results. Where the training datasets are Astral25 and Astral40 and testing datasets are 25PDB and 1189, respectively, the results obtained are similar to 10-fold and leave-one-out but with a lower standard deviation. Results show that using larger training datasets provide more information (a robust predictive model) to predict the testing dataset. To end, the better test procedure is independent-sets where the training datasets is larger and independent of the testing datasets.

7.6 Assessment of multiple k-nearest neighbour

As described in chapter 3, multiple k-nearest neighbour (MKNN) classifiers extends KNN by combining different *k*'s to achieve a better result than using a single K. An observation was made with the selection of combined K's when the independent-set test procedure was used, where the testing datasets are 25PDB and 1189, in most of the analyses, particularly the top 10 results as shown in Table 4-13 and Table 5-11 the KNN value of the predicted class label was consistently at 1. This is the result of when a large dataset (i.e., Astral25 & Astral40) are used as training datasets, as these datasets are larger than the testing datasets there is much more information for a test sample to be accurately classified. In the nearest neighbour model, the smaller datasets use just over the midpoint number of KNN, and the larger dataset size ones are using the higher *k*, which is attributed due to the datasets large sample size.

Each of the 611 amino acid indices resulted in many tens of thousands of individual results, e.g. AAI # 1 ANDN920101 generated 2047 rows of results over 10 folds per class which equates to (2^11)-1, plus the average (mean) over the 4 classes, in total 10235(2047*5) individual

accuracies are obtained. Leave-one-out and independent sets generated (2^11)-1 (2047 models). Results presented in Table 7-1, Table 7-2 and Table 7-3 are derived from the mean set of accuracies. The interesting output from the k-model point of view is the higher the frequency of a particular AAI # the likely that the index has the highest accuracy over 2047 models which correlates with the top 10 results presented in Table 5-15. The count column is the frequency of an index that appears with the highest accuracy over 2047 models. The indices presented in Table 7-1, Table 7-2 and Table 7-3 results are consistent with the range of top 10 indices across all datasets and test procedure combinations.

Table 7-1 Assessment of k neighbours using 10-fold test procedure

Rank	25PDB	Count	1189	Count	Astral25	Count	Astral40	Count
1	**414**	2032	**437**	1921	160	935	**568**	1257
2	415	12	414	38	409	444	160	770
3	495	3	467	38	230	326	64	12
4			154	15	414	132	412	7
5			97	12	568	130	414	1
6			143	11	437	39		
7			466	4	346	21		
8			407	3	162	7		
9			463	2	532	6		
10			468	2	466	4		
11			134	1	64	1		
12					137	1		
13					170	1		

Table 7-2 Assessment of k neighbours using LOO test procedure

Rank	25PDB	Count	1189	Count	Astral25	Count	Astral40	Count
1	**414**	2031	**437**	1344	**466**	747	64	1020
2	415	15	463	103	160	537	437	410
3	343	1	414	99	437	375	31	187
4			328	16	532	239	532	133
5			72	12	409	72	160	116
6			262	10	136	66	134	70
7			466	9	64	6	468	35
8			464	8	137	3	136	30
9			302	5	568	2	454	15
10			31	3			568	12
11			154	3			456	8
12			411	3			414	7
13			143	2			230	2
14			438	2			345	1
15			75	1			466	1
16			346	1				
17			409	1				
18								

Table 7-3 Assessment of k neighbours using independent-sets test procedure

Rank	25PDB	Count	1189	Count	Astral25	Count	Astral40	Count
1	466	1216	302	1860	**414**	1884	414	1105
2	31	249	31	140	573	93	568	930
3	188	211	437	40	31	27	29	12
4	467	201	136	3	468	24		
5	64	86	28	2	581	17		
6	134	54	137	1	350	2		
7	469	12	198	1				
8	456	7						
9	136	4						
10	454	3						
11	137	2						
12	196	1						
13	414	1						

7.7 Conclusions

Prediction of protein structural classes is an important area in proteomics because (1) enables the identification of common structural patterns in proteins (2) reduces the conformational search space during the search of the tertiary structure and (3) knowledge of structural classes is a useful information applicable to the wider area of proteomics. The comprehensive investigations into traditional sequence driven features was undertaken to analyse and evaluate the effects of different factors, which are, different datasets, large sample size datasets, different homology levels and different test procedures. The four datasets used are 25PDB, 1189, Astral25 and Astral40. Two of which are standard datasets used in many structural class studies (25PDB and 1189) and the other two datasets are largest protein structural classes ever to be dataset constructed, Astral25 and Astral40. Two different homology levels (25% and 40%) were considered when selecting the datasets, the 25% homology level are 25PDB and Astral25, the 40% homology levels datasets are 1189 and Astral40. The three different test procedures are n-fold, leave one out and independent-sets. The reason why many different factors were considered was that no existing study had looked into the effects of these multiple factors; studies tend to look into using a single dataset, single homology level and a single test procedure. The most important reason for undertaking the investigation was to analyse the largest set of traditional sequence driven features to determine which features or groups of features are better suited for the prediction of protein structural classes. The results show that the analyses over different dataset, homology levels and test procedures produce varying results and that no one single method should be relied upon instead a consensus of decisions and results should be taking into consideration to build a reliable model. The results also show that the best type of features for the prediction of

protein structural classes are composition based such as amino acid composition and dipeptide composition. Chapter 5 presented four methods utilising amino acid indices. It is within the utilisation of the amino acid indices where success has been shown towards the prediction of protein structural classes. Similarly to chapter 4 large-scale analysis into traditional sequence features a range of analyses was applied to the largest set of amino acid indices dataset – which was developed within this thesis and contained 611 of them. There had previously been no large-scale analysis over amino acid indices, as was evident with findings in chapter four that several feature groups, namely autocorrelation and PseAAC utilised a very limited number of the available amino acid indices. This analysis had identified five amino acid indices that are the best candidates towards prediction of protein structural classes, which are amino acid index 414 (AURR980115), index 31 (CHAM830108), index 437 (KUMS000101), index 466 (FUKS010109), and acid index 568 – average medium contacts (N_m). The proposed hybrid system has helped not only refine the current amino acid indices list but also generate new amino acid indices that have been successfully shown to better characterise the proteins compared to the existing amino acid indices. Results show that the computationally generated indices have higher predictive accuracies compared to the individual amino acid indices it clustered. The application of the hybrid method is more significant as it shows that clustering similar indices removes redundancy and in its place summarises the clustered indices with a single index that represents 100% of the original data variability of the original group of amino acid indices, which characterise structural classes of proteins far better. The SRM is a novel way to represent each protein sample in over 600 ways – i.e. each protein sample is represented by each amino acid index in one vector i.e. the mean (method 3) or PCA (method 4). The use of the mean and PCA is a statistical way to summarise the SRM. Feature selection methods was applied over the traditional sequence driven feature space and method 3 and method 4 feature spaces, predictive accuracy higher than the hybrid or the GAAC method was not significant with the selection feature. However, the feature selection methods highlighted the most representative set of features, which are the composition based secondary structure (feature index 6.1 and 6.1.6), which linked with the top 10 results from Table 4-13 in chapter 4. The performance of our approaches compared to traditional sequence-based methods had an overall accuracy of 75.52% using the 25PDB dataset, which compared to the results presented in Table 2-8 is the highest. The goal of this thesis was to identify and extract new sequence driven features for the prediction of protein structural classes. For this purpose, a novel

approach that utilises amino acid indices was achieved. Below is a summary of conclusion and contributions.

- Composition based feature groups are the most useful descriptors at predicting structural classes of proteins. Further developments should be made using composition based features and amino acid indices.
- Hydrophobicity and secondary structure feature indices are found to be the most important physicochemical properties of amino acids and are generally found to be more related towards structural classes of proteins than another property.
- PseAAC feature group's predictive power comes from the amino acid composition aspect – the study carried out in chapter 4 shows that the sequence-order-effect (lambda) did not yield any significant improvement in accuracy using the benched mark datasets and the larger Astral datasets towards protein structural classes. The limited selection of amino acid indices could have been expanded to the full dataset.
- The further advancement and refinement of the amino acid indices, has further extended to include as many new amino acid indices found in literature.
- Generalising the amino acid composition to take into account the many hundred natural weights of amino acids – this concept can be applied to any feature that utilises amino acid indices. Compared with other approaches, the GAAC method is effective and powerful in improving the overall predictive accuracy and is only limited to the current number of available amino acid indices. In this thesis, a newly designed generalised feature set combined with existing and newly found amino acid indices are employed to predict structural classes for low similarity protein sequences. The resulted feature vectors, each representing one protein, fed into the MKNN algorithm for the prediction of protein structural classes. Comparing with benched marked results the GAAC method presented higher prediction accuracy in all cases. In addition, it is shown that careful selection of natural amino acid values are important to achieve good prediction accuracies for whatever proteomic characteristics being predicted.
- New computationally derived amino acid indices had better represent subset of the original set of amino acid indices – these indices can be found at http://cisaps.com. Hybrid sequence driven feature extraction method summarised the amino acid index dataset so well it produces one of the best predictive accuracy for 25% homology

datasets 25PDB. These new computationally derived amino acid indices can replace or be used alongside the original amino acid dataset

- The development of the GAAC webserver, which is freely accessible to bioinformaticians at http://www.generalised-protein-sequence-features.com/.

The work presented in this thesis has further expanded the understanding and knowledge of sequence driven features in the context of predicting protein structural classes by the development of feature extraction methods utilising amino acid indices.

7.8 Future work

This section highlights possible future works towards prediction of structural classes of proteins and other proteomics areas.

- Chapter 5 presented two feature extractions method, methods 3 and 4, which uses the SRM to represent each protein sequence with either the mean (method 3) or PCA (method 4) statistical properties. The next statistical property to be investigated is independent component analysis (ICA). The application of ICA over the SRM can be used to derive ICA to generated computationally derived amino acid indices.
- The generalisation approach developed for the GAAC method can be implemented for the autocorrelation and PseAAC feature groups, which currently utilises a limited number of amino acid indices. Generalisation of these feature groups will allow the full utilisation of the updated amino acid indices dataset and the computationally generated amino acid indices, and results generated from using these indices may increase further.
- Search for a better set of training samples to better categorise the testing data, early experimental analysis suggestion that searching for a representative subset of protein samples from the training dataset increase the overall predictive accuracy.
- Exploring the use of the feature extraction methods presented in this thesis in other proteomic such as prediction of protein melting points, protein to protein interaction, protein expressions, protein subcellular prediction, to name but a few, practically the feature extraction methods can be used in any area where the raw data are the proteins primary sequence may yield better results.

References

A. Brazma, H. P., T. Schlitt and M Shojatalab. (2002). "Basic introduction to molecular biology." from http://www.ebi.ac.uk/2can/biology/.

Ahmadi Adl, A., A. Nowzari-Dalini, et al. (2012). "Accurate prediction of protein structural classes using functional domains and predicted secondary structure sequences." Journal of Biomolecular Structure & Dynamics **29**(6): 623-33.

Aizerman, A., E. M. Braverman, et al. (1964). "Theoretical foundations of the potential function method in pattern recognition learning." Automation and Remote Control **25**: 821-837.

Anand, A., G. Pugalenthi, et al. (2008). "Predicting protein structural class by SVM with class-wise optimized features and decision probabilities." Journal of Theoretical Biology **In Press, Corrected Proof**.

Andersen, N. H., B. L. Cao, et al. (1992). "PEPTIDE PROTEIN-STRUCTURE ANALYSIS USING THE CHEMICAL-SHIFT INDEX METHOD - UPFIELD ALPHA-CH VALUES REVEAL DYNAMIC HELICES AND ALPHA-L SITES." Biochemical and Biophysical Research Communications **184**(2): 1008-1014.

Asakawa, N., N. Sakiyama, et al. (2010). "Characteristic amino acid distribution around segments unique to allergens." Journal of Biochemistry **147**(1): 127-133.

Atchley, W. R., J. Zhao, et al. (2005). "Solving the protein sequence metric problem." Proceedings of the National Academy of Sciences of the United States of America **102**(18): 6395-6400.

Aurora, R. and G. D. Rose (1998). "Helix capping." Protein Science **7**(1): 21-38.

Bahar, I., A. R. Atilgan, et al. (1997). "Understanding the recognition of protein structural classes by amino acid composition." Proteins: Structure, Function and Genetics **29**(2): 172-185.

Berman, H. M. (2007). "The Protein Data Bank: A historical perspective." Acta Crystallographica Section A: Foundations of Crystallography **64**(1): 88-95.

Bernstein, F. C., T. F. Koetzle, et al. (1977). "PROTEIN DATA BANK - COMPUTER-BASED ARCHIVAL FILE FOR MACROMOLECULAR STRUCTURES." Journal of Molecular Biology **112**(3): 535-542.

Bhasin, M. and G. P. S. Raghava (2004). "Classification of nuclear receptors based on amino acid composition and dipeptide composition." Journal of Biological Chemistry **279**(22): 23262-23266.

Bhaskaran, R. and P. K. Ponnuswamy (1988). "POSITIONAL FLEXIBILITIES OF AMINO-ACID RESIDUES IN GLOBULAR-PROTEINS." International Journal of Peptide and Protein Research **32**(4): 241-255.

Bigelow, C. C. (1967). "On the average hydrophobicity of proteins and the relation between it and protein structure." Journal of Theoretical Biology **16**(2): 187-211.

Bock, J. R. and D. A. Gough (2001). "Predicting protein-protein interactions from primary structure." Bioinformatics **17**(5): 455-460.

Boser, B., I. Guyon, et al. (1992). A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory, Pittsburgh, Pennsylvania, United States, ACM.

Broto, P., G. Moreau, et al. (1984). "MOLECULAR-STRUCTURES - PERCEPTION, AUTO-CORRELATION DESCRIPTOR AND SAR STUDIES - SYSTEM OF ATOMIC CONTRIBUTIONS FOR THE CALCULATION OF THE NORMAL-OCTANOL WATER PARTITION-COEFFICIENTS." European Journal of Medicinal Chemistry **19**(1): 71-78.

Bu, W. S., Z. P. Feng, et al. (1999). "Prediction of protein (domain) structural classes based on amino-acid index." European Journal of Biochemistry **266**(3): 1043-1049.

Byvatov, E. and G. Schneider (2003). "Support vector machine applications in bioinformatics." Applied bioinformatics **2**(2): 67-77.

Cai, C. Z., L. Y. Han, et al. (2003). "SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence." Nucleic Acids Research **31**(13): 3692-3697.

Cai, Y.-D., X.-J. Liu, et al. (2001). "Support Vector Machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect." Journal of Cellular Biochemistry **84**(2): 343-348.

Cai, Y. D., X. J. Liu, et al. (2001). "Support vector machines for predicting protein structural class." BMC bioinformatics [electronic resource] **2**(1): 3.

Cai, Y. D., X. J. Liu, et al. (2002). "Prediction of protein structural classes by support vector machines." Computers & Chemistry **26**(3): 293-296.

Cai, Y. D., X. J. Liu, et al. (2003). "Support vector machines for prediction of protein domain structural class." Journal of Theoretical Biology **221**(1): 115-120.

Chandonia, J. M., G. Hon, et al. (2004). "The ASTRAL Compendium in 2004." Nucleic Acids Research **32**(DATABASE ISS.).

Chang, C.-C. and C.-J. Lin (2001). "LIBSVM: a library for support vector machines."

Charton, M. (1981). "PROTEIN FOLDING AND THE GENETIC-CODE - AN ALTERNATIVE QUANTITATIVE MODEL." Journal of Theoretical Biology **91**(1): 115-123.

Charton, M. and B. I. Charton (1982). "THE STRUCTURAL DEPENDENCE OF AMINO-ACID HYDROPHOBICITY PARAMETERS." Journal of Theoretical Biology **99**(4): 629-644.

Charton, M. and B. I. Charton (1983). "THE DEPENDENCE OF THE CHOU-FASMAN PARAMETERS ON AMINO-ACID SIDE-CHAIN STRUCTURE." Journal of Theoretical Biology **102**(1): 121-134.

Chen, C., L. X. Chen, et al. (2008). "Predicting protein structural class based on multi-features fusion." Journal of Theoretical Biology **253**(2): 388-392.

Chen, C., Y.-X. Tian, et al. (2006). "Using pseudo-amino acid composition and support vector machine to predict protein structural class." Journal of Theoretical Biology **243**(3): 444-448.

Chen, C., X. Zhou, et al. (2006). "Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network." Analytical Biochemistry **357**(1): 116-121.

Chen, C. H. (1978). "Pattern recognition and signal processing." Problemy Sluchainogo Poiska.

Chen, L., L. Lu, et al. (2009). "Multiple classifier integration for the prediction of protein structural classes." Journal of Computational Chemistry **30**(14): 2248-2254.

Chen, W., S. W. Zhang, et al. (2008). Prediction of seven protein structural classes by fusing multi-feature information including protein evolutionary conservation information. 2nd International Conference on Bioinformatics and Biomedical Engineering, iCBBE 2008.

Chen, Y., F. Chen, et al. (2008). "Ensemble voting system for multiclass protein fold recognition." International Journal of Pattern Recognition and Artificial Intelligence **22**(4): 747-763.

Chothia, C. (1976). "The nature of the accessible and buried surfaces in proteins." Journal of Molecular Biology **105**(1): 1-12.

Chou, K.-C. (1999). "A Key Driving Force in Determination of Protein Structural Classes." Biochemical and Biophysical Research Communications **264**(1): 216-224.

Chou, K. C. (1995). "A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space." Proteins: Structure, Function and Genetics **21**(4): 319-344.

Chou, K. C. (2001). "Prediction of protein cellular attributes using pseudo-amino acid composition." Proteins: Structure, Function and Genetics **43**(3): 246-255.
Chou, K. C. (2004). "Structural bioinformatics and its impact to biomedical science." Current Medicinal Chemistry **11**(16): 2105-2134.
Chou, K. C. (2005). "Progress in protein structural class prediction and its impact to bioinformatics and proteomics." Current Protein and Peptide Science **6**(5): 423-436.
Chou, K. C. (2009). "Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology." Current Proteomics **6**(4): 262-274.
Chou, K. C. (2011). "Some remarks on protein attribute prediction and pseudo amino acid composition." Journal of Theoretical Biology **273**(1): 236-247.
Chou, K. C. and Y. D. Cai (2004). "Predicting protein structural class by functional domain composition." Biochemical and Biophysical Research Communications **321**(4): 1007-1009.
Chou, K. C. and Y. D. Cai (2004). "Prediction of protein subcellular locations by GO-FunD-PseAA predictor." Biochemical and Biophysical Research Communications **320**(4): 1236-1239.
Chou, K. C. and G. M. Maggiora (1998). "Domain structural class prediction." Protein Engineering **11**(7): 523-538.
Chou, K. C. and C. T. Zhang (1995). "Prediction of protein structural classes." Critical Reviews in Biochemistry and Molecular Biology **30**(4): 275-349.
Chou, P. Y. (1989). "Prediction of protein structural classes from amino acid composition." Prediction of Protein Structure and the Principles of Protein Conformation: 549-586.
Cid, H., M. Bunster, et al. (1992). "Hydrophobicity and structural classes in proteins." Protein Engineering **5**(5): 373-375.
Cohen, F. E. and I. D. Kuntz (1987). "PREDICTION OF THE 3-DIMENSIONAL STRUCTURE OF HUMAN GROWTH-HORMONE." Proteins-Structure Function and Genetics **2**(2): 162-166.
Cohen, F. E. and I. D. Kuntz (1987). "Prediction of the three-dimensional structure of human growth hormone." Proteins: Structure, Function and Genetics **2**(2): 162-166.
Cohen, I., Q. Tian, et al. (2002). "Feature selection using principal feature analysis." Univ. of Illinois at Urbana-Champaign.
Consortium, T. U. (2012). "Reorganizing the protein space at the Universal Protein Resource (UniProt)." Nucleic Acids Research **40**(Database issue): D71-5.
Cortes, C. and V. Vapnik (1995). "SUPPORT-VECTOR NETWORKS." Machine Learning **20**(3): 273-297.
Costantini, S., G. Colonna, et al. (2007). "PreSSAPro: A software for the prediction of secondary structure by amino acid properties." Computational Biology and Chemistry **31**(5-6): 389-392.
Costantini, S. and A. M. Facchiano (2008). "Prediction of the protein structural class by specific peptide frequencies." Biochimie **In Press, Corrected Proof**.
Cover, T. and P. Hart (1967). "Nearest neighbor pattern classification." Information Theory, IEEE Transactions on **13**(1): 21-27.
Csaba, G., F. Birzele, et al. (2009). "Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis." BMC Structural Biology **9**(1): 23.
Cui, J., L. Y. Han, et al. (2007). "Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties." Molecular Immunology **44**(5): 866-877.
Davies, M. N., A. Secker, et al. (2007). "On the hierarchical classification of G protein-coupled receptors." Bioinformatics **23**: 3113 - 3118.

Davis, G. J., W. F. Bosron, et al. (1996). "X-ray structure of human beta(3)beta(3) alcohol dehydrogenase - The contribution of ionic interactions to coenzyme binding." Journal of Biological Chemistry **271**(29): 17057-17061.

Dayhoff, H., Calderone,H (1978). "Composition of Proteins." Altas of Protein Sequence and Structure **5**: 363–373.

Dayhoff, M. O., R. V. Eck, et al. (1972). A MODEL OF EVOLUTIONARY CHANGE IN PROTEINS.

Deleage, G. and J. S. Dixon (1989). "Use of class prediction to improve protein secondary structure prediction." Prediction of Protein Structure and the Principles of Protein Conformation: 587-597.

Deleage, G. and B. Roux (1987). "An algorithm for protein secondary structure prediction based on class prediction." Protein Engineering **1**(4): 289-294.

Ding, C., H. C. Peng, et al. (2003). Minimum redundancy feature selection from microarray gene expression data.

Ding, C. H. Q. and I. Dubchak (2001). "Multi-class protein fold recognition using support vector machines and neural networks." Bioinformatics **17**(4): 349-358.

Ding, S., S. Zhang, et al. (2012). "A novel protein structural classes prediction method based on predicted secondary structure." Biochimie **94**(5): 1166-1171.

Ding, Y. S., T. L. Zhang, et al. (2007). "Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network." Protein and Peptide Letters **14**(8): 811-815.

Ding, Y. S., T. L. Zhang, et al. (2009). "Using maximum entropy model to predict protein secondary structure with single sequence." Protein and Peptide Letters **16**(5): 552-560.

Du, Q. S., Z. Q. Jiang, et al. (2006). "Amino Acid Principal Component Analysis (AAPCA) and its applications in protein structural class prediction." Journal of Biomolecular Structure & Dynamics **23**(6): 635-640.

Dubchak, I., I. Muchink, et al. (1995). "Prediction of protein folding class using global description of amino acid sequence." Proc Natl Acad Sci USA **92**: 8700 - 8704.

Dubchak, I., I. Muchnik, et al. (1999). "Recognition of a protein fold in the context of the SCOP classification." Proteins-Structure Function and Genetics **35**(4): 401-407.

Eidhammer I, J. I., Taylor W.R (2005). "Protein Bioinformatics."

Eisenberg, D. (2003). "The discovery of the α-helix and β-sheet, the principal structural features of proteins." Proceedings of the National Academy of Sciences of the United States of America **100**(20): 11207-11210.

Eisenhaber, F., C. Frömmel, et al. (1996). "Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class." Proteins: Structure, Function and Genetics **25**(2): 169-179.

Eisenhaber, F., F. Imperiale, et al. (1996). "Prediction of secondary structural content of proteins from their amino acid composition alone. I. New analytic vector decomposition methods." Proteins: Structure, Function and Genetics **25**(2): 157-168.

Feitelson, D. G. and M. Treinin (2002). "The blueprint for life." Computer **35**(7).

Fernandez, L., J. Caballero, et al. (2007). "Amino acid sequence autocorrelation vectors and Bayesian-regularized genetic neural networks for modeling protein conformational stability: Gene V protein mutants." Proteins: Structure, Function and Genetics **67**(4): 834-852.

Fukuchi, S. and K. Nishikawa (2001). "Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria." Journal of Molecular Biology **309**(4): 835-843.

Georgiev, A. G. (2009). "Interpretable numerical descriptors of amino acid space." Journal of Computational Biology **16**(5): 703-723.

Goldstein, P., J. Zucko, et al. (2009). "Clustering of protein domains for functional and evolutionary studies." BMC Bioinformatics **10**: 335.

Goodman, S. and A. Hunter (1999). Feature extraction algorithms for pattern classification. IEE Conference Publication.

Goodsell, S. D. (2010). "The Protein Data Bank: Exploring Biomolecular Structure." Nature Education.

Gorga, F. K. (2008). "Introduction to Protein Structure." Bridgewater State College, MA.

Grassmann, J., M. Reczko, et al. (1999). "Protein fold class prediction: new methods of statistical classification." Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology: 106-12.

Grassmann, J., M. Reczko, et al. (1999). "Protein fold class prediction: new methods of statistical classification." Proceedings / . International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology: 106-112.

Gromiha, M. M. and S. Selvaraj (1998). "Protein secondary structure prediction in different structural classes." Protein Engineering **11**(4): 249-251.

Gromiha, M. M. and S. Selvaraj (2004). "Inter-residue interactions in protein folding and stability." Progress in Biophysics & Molecular Biology **86**(2): 235-277.

Gu, F. and H. Chen (2009). "Evaluating long-term relationship of protein sequence by use of D-interval conditional probability and its impact on protein structural class prediction." Protein and Peptide Letters **16**(10): 1267-1276.

Han, P., X. Zhang, et al. (2009). "Predicting disordered regions in proteins using the profiles of amino acid indices." BMC Bioinformatics **10**(SUPPL. 1).

Hansen, J. C., X. Lu, et al. (2006). "Intrinsic protein disorder, amino acid composition, and histone terminal domains." Journal of Biological Chemistry **281**(4): 1853-1856.

Hernández-Rodríguez, S., J. F. Martínez-Trinidad, et al. (2010). "Fast k most similar neighbor classifier for mixed data (tree k-MSN)." Pattern Recognition **43**(3): 873-886.

Hmeidi, I., B. Hawashin, et al. (2008). "Performance of KNN and SVM classifiers on full word Arabic articles." Advanced Engineering Informatics **22**(1): 106-111.

Hobohm, U. and C. Sander (1994). "Enlarged Representative Set Of Protein Structures." Protein Science **3**(3): 522-524.

Hooker, C. A. (1995). "ADAPTATION IN NATURAL AND ARTIFICIAL SYSTEMS - HOLLAND,JH." Philosophical Psychology **8**(3): 287-299.

Horne, D. S. (1988). "PREDICTION OF PROTEIN HELIX CONTENT FROM AN AUTO-CORRELATION ANALYSIS OF SEQUENCE HYDROPHOBICITIES." Biopolymers **27**(3): 451-477.

Hotta, S., S. Kiyasu, et al. (2004). A classifier based on distance between test samples and average patterns of categorical nearest neighbors. Proceedings - International Workshop on Frontiers in Handwriting Recognition, IWFHR.

Hua, J., W. D. Tembe, et al. (2009). "Performance of feature-selection methods in the classification of high-dimension data." Pattern Recognition **42**(3): 409-424.

Huang, J., S. Kawashima, et al. (2007). "New amino acid indices based on residue network topology." Genome informatics. International Conference on Genome Informatics **18**: 152-161.

Isik, Z., B. Yanikoglu, et al. (2004). Protein structural class determination using support vector machines. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). **3280:** 82-89.

Jahandideh, S., S. Hoseini, et al. (2009). "A hybrid genetic-neural model for predicting protein structural classes." Biologia **64**(4): 649-654.

Jayaraman, V. K. and V. Sundararajan (2010). Applications of Support Vector Machines In Chemo And Bioinformatics. International Conference on Modeling, Optimization, and Computing. S. Paruya, S. Kar and S. Roy. **1298:** 18-23.

Jian-Ding, Q. I. U., L. U. O. San-Hua, et al. (2009). "Using support vector machines for prediction of protein structural classes based on discrete wavelet transform." Journal of Computational Chemistry **30**(8): 1344-1350.

Jin, L., W. Fang, et al. (2003). "Prediction of protein structural classes by a new measure of information discrepancy." Computational Biology and Chemistry **27**(3): 373-380.

Karadaghi, S. A. (2012). "Introduction to protein structure and structural bioinformatics."

Karchin, R., K. Karplus, et al. (2002). "Classifying G-protein coupled receptors with support vector machines." Bioinformatics **18**: 147 - 159.

Kawashima, S. and M. Kanehisa (2000). "AAindex: Amino acid index database." Nucleic Acids Research **28**(1): 374-374.

Kawashima, S., P. Pokarowski, et al. (2008). "AAindex: Amino acid index database, progress report 2008." Nucleic Acids Research **36**(SUPPL. 1).

Kazemian, M., B. Moshiri, et al. (2007). "A new expertness index for assessment of secondary structure prediction engines." Computational Biology and Chemistry **31**(1): 44-47.

Ke Chen, L. A. K. J. R. (2008). "Prediction of protein structural class using novel evolutionary collocation-based sequence representation." Journal of Computational Chemistry **9999**(9999): NA.

Kedarisetti, K. D., L. Kurgan, et al. (2006). "Classifier ensembles for protein structural class prediction with varying homology." Biochemical and Biophysical Research Communications **348**(3): 981-988.

Kertész-Farkas, A., S. Dhir, et al. (2007). "Benchmarking protein classification algorithms via supervised cross-validation." Journal of Biochemical and Biophysical Methods **In Press, Corrected Proof**.

Kidera, A., Y. Konishi, et al. (1985). "STATISTICAL-ANALYSIS OF THE PHYSICAL-PROPERTIES OF THE 20 NATURALLY-OCCURRING AMINO-ACIDS." Journal of Protein Chemistry **4**(1): 23-55.

Kneller, D. G., F. E. Cohen, et al. (1990). "Improvements in protein secondary structure prediction by an enhanced neural network." Journal of Molecular Biology **214**(1): 171-182.

Kumar, S., C. J. Tsai, et al. (2000). "Factors enhancing protein thermostability." Protein Engineering **13**(3): 179-191.

Kumarevel, T. S., M. M. Gromiha, et al. (2000). "Structural class prediction: an application of residue distribution along the sequence." Biophysical Chemistry **88**(1-3): 81-101.

Kurgan, L. and K. Chen (2007). "Prediction of protein structural class for the twilight zone sequences." Biochemical and Biophysical Research Communications **357**(2): 453-460.

Kurgan, L., K. Cios, et al. (2008). "SCPRED: Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences." BMC Bioinformatics **9**(1): 226.

Kurgan, L. A. and L. Homaeian (2006). "Prediction of structural classes for protein sequences and domains-Impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy." Pattern Recognition **39**(12): 2323-2343.

Kurgan, L. A., W. Stach, et al. (2007). "Novel scales based on hydrophobicity indices for secondary protein structure." Journal of Theoretical Biology **248**(2): 354-366.

Kyoung Kim, J., S.-Y. Bang, et al. (2006). "Sequence-driven features for prediction of subcellular localization of proteins." Pattern Recognition **39**(12): 2301-2311.

Lamond, A. I. (2002). "Molecular biology of the cell, 4th edition." Nature **417**(6887): 383-383.

Levitt, M. and C. Chothia (1976). "Structural patterns in globular proteins." Nature **261**(5561): 552-558.
Li, B.-Q., T. Huang, et al. (2012). "Identification of Colorectal Cancer Related Genes with mRMR and Shortest Path in Protein-Protein Interaction Network." Plos One **7**(3): e33393.
Li, B. Q., L. L. Hu, et al. (2012). "Prediction of protein domain with mRMR feature selection and analysis." Plos One **7**(6).
Li, W., K. Lin, et al. (2008). "Prediction of protein structural classes using hybrid properties." Molecular Diversity **12**(3-4): 171-179.
Li, W. Z., L. Jaroszewski, et al. (2001). "Clustering of highly homologous sequences to reduce the size of large protein databases." Bioinformatics **17**(3): 282-283.
Li, W. Z., L. Jaroszewski, et al. (2002). "Tolerating some redundancy significantly speeds up clustering of large protein databases." Bioinformatics **18**(1): 77-82.
Li, Z. C., X. B. Zhou, et al. (2008). "Prediction of protein structure class by coupling improved genetic algorithm and support vector machine." Amino Acids **35**(3): 581-590.
Li, Z. R., H. H. Lin, et al. (2006). "PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence." Nucleic Acids Res **34**: W32 - W37.
Li, Z. R., H. H. Lin, et al. (2006). "PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence." Nucleic Acids Research **34**(WEB. SERV. ISS.).
Liang, Z. and T. Zhao (2006). Feature selection for linear support vector machines. Proceedings - International Conference on Pattern Recognition.
Lin, H. and Q. Z. Li (2007). "Using pseudo amino acid composition to predict protein structural class: Approached by incorporating 400 dipeptide components." Journal of Computational Chemistry **28**(9): 1463-1466.
Lin, K. L., L. Chun-Yuan, et al. (2007). "Feature Selection and Combination Criteria for Improving Accuracy in Protein Structure Prediction." NanoBioscience, IEEE Transactions on **6**(2): 186-196.
Lin, Y.-W. C. a. C.-J. (2005). "Combining SVMs with Various Feature Selection Strategies."
Lin, Z. and X. M. Pan (2001). "Accurate prediction of protein secondary structural content." Journal of Protein Chemistry **20**(3): 217-220.
Liu, T., X. Zheng, et al. (2009). "Prediction of protein structural class using a complexity-based distance measure." Amino Acids: 1-8.
Lo, S. L., C. Z. Cai, et al. (2005). "Effect of training datasets on support vector machine prediction of protein-protein interactions." Proteomics **5**(4): 876-884.
Lopes, F. M., D. C. Martins Jr, et al. (2008). "Feature selection environment for genomic applications." BMC Bioinformatics **9**.
Luo, R. Y., Z. P. Feng, et al. (2002). "Prediction of protein structural class by amino acid and polypeptide composition." European Journal of Biochemistry **269**(17): 4219-4225.
Ma, S. and Y. Dai (2011). "Principal component analysis based methods in bioinformatics studies." Briefings in Bioinformatics **12**(6): 714-722.
Ma, X., J. S. Wu, et al. (2009). A SVM-based approach for predicting DNA-binding residues in proteins from amino acid sequences.
Markowetz, F., L. Edler, et al. (2003). "Support vector machines for protein fold class prediction." Biometrical Journal **45**(3): 377-389.
Marsolo, K. and S. Parthasarathy (2006). On the Use of Structure and Sequence-Based Features for Protein Classification and Retrieval. Data Mining, 2006. ICDM '06. Sixth International Conference on.
Marx, M. L. a. L., R.J.} (2006). "Introduction to mathematical statistics and its applications},."

Melvin, I., J. Weston, et al. (2008). "Combining classifiers for improved classification of proteins from sequence or structure." BMC Bioinformatics **9**.

Metfessel, B. A., P. N. Saurugger, et al. (1993). "Cross-validation of protein structural class prediction using statistical clustering and neural networks." Protein Science **2**(7): 1171-1182.

Mielke, S. P. and V. V. Krishnan (2003). "Protein structural class identification directly from NMR spectra using averaged chemical shifts." Bioinformatics **19**(16): 2054-2064.

Mizianty, M. and L. Kurgan (2009). "Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences." BMC Bioinformatics **10**(1): 414.

Moll, M. and L. E. Kavraki (2008). "Matching of structural motifs using hashing on residue labels and geometric filtering for protein function prediction." Comput Syst Bioinformatics Conf **7**: 157-68.

Muggleton, S., R. D. King, et al. (1992). "Protein secondary structure prediction using logic-based machine learning." Protein Engineering **5**(7): 647-657.

Murzin, A. G., S. E. Brenner, et al. (1995). "SCOP: A structural classification of proteins database for the investigation of sequences and structures." Journal of Molecular Biology **247**(4): 536-540.

Nakai, K., A. Kidera, et al. (1988). "CLUSTER-ANALYSIS OF AMINO-ACID INDEXES FOR PREDICTION OF PROTEIN-STRUCTURE AND FUNCTION." Protein Engineering **2**(2): 93-100.

Nakashima, H., K. Nishikawa, et al. (1986). "The folding type of a protein is relevant to the amino acid composition." Journal of Biochemistry **99**(1): 153-162.

Nanuwa, S. S., A. Dziurla, et al. (2009). Weighted amino acid composition based on amino acid indices for prediction of protein structural classes. Information Technology and Applications in Biomedicine, 2009. ITAB 2009. 9th International Conference on.

Nanuwa, S. S. and H. Seker (2008). Investigation into the role of sequence-driven-features for prediction of protein structural classes. BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on.

Ni, B. and J. Liu (2004). A hybrid filter/wrapper gene selection method for microarray classification. Proceedings of 2004 International Conference on Machine Learning and Cybernetics.

Nigsch, F., A. Bender, et al. (2006). "Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization." Journal of Chemical Information and Modeling **46**(6): 2412-2422.

Nishikawa, K. and T. Ooi (1980). "Prediction of the surface-interior diagram of globular proteins by an empirical method." International Journal of Peptide and Protein Research **16**(1): 19-32.

Nishikawa, K. and T. Ooi (1982). "CORRELATION OF THE AMINO-ACID-COMPOSITION OF A PROTEIN TO ITS STRUCTURAL AND BIOLOGICAL CHARACTERS." Journal of Biochemistry **91**(5): 1821-1824.

Nishikawa, K. and T. Ooi (1986). "Radial locations of amino acid residues in a globular protein: Correlation with the sequence." Journal of Biochemistry **100**(4): 1043-1047.

Ong, S., H. Lin, et al. (2007). "Efficacy of different protein descriptors in predicting protein functional families." BMC Bioinformatics **8**(1): 300.

Orengo, C. A., A. D. Michie, et al. (1997). "CATH - a hierarchic classification of protein domain structures." Structure **5**(8): 1093-1108.

Overington, J. P., B. Al-Lazikani, et al. (2006). "How many drug targets are there?" Nat Rev Drug Discov **5**(12): 993-996.

Peng, H. C., F. H. Long, et al. (2005). "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy." Ieee Transactions on Pattern Analysis and Machine Intelligence **27**(8): 1226-1238.

Popov, I., A. Nenov, et al. (2009). "BIOINFORMATICS IN PROTEOMICS: A REVIEW ON METHODS AND ALGORITHMS." Biotechnology & Biotechnological Equipment **23**(1): 1115-1120.

Pruitt, K. D., T. Tatusova, et al. (2012). "NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy." Nucleic Acids Research **40**(D1): D130-D135.

Pruitt, K. D., T. Tatusova, et al. (2009). "NCBI Reference Sequences: current status, policy and new initiatives." Nucleic Acids Research **37**: D32-D36.

Purves, W. K. (2001). Life: The Science of Biology.

Querol, E., J. A. PerezPons, et al. (1996). "Analysis of protein conformational characteristics related to thermostability." Protein Engineering **9**(3): 265-271.

Reczko, M. and H. Bohr (1994). "THE DEF DATA-BASE OF SEQUENCE BASED PROTEIN FOLD CLASS PREDICTIONS." Nucleic Acids Research **22**(17): 3616-3619.

Ron, K. and H. J. George (1997). "Wrappers for feature subset selection." Artif. Intell. **97**(1-2): 273-324.

Rost, B. (1998). "Protein Structure Prediction in 1D, 2D and 3D." European Molecular Biology Laboratory.

Rost, B. (1999). "Twilight zone of protein sequence alignments." Protein Engineering **12**(2): 85-94.

Roy, S., D. Martinez, et al. (2009). "Exploiting Amino Acid Composition for Predicting Protein-Protein Interactions." Plos One **4**(11).

Russell, S. (2012). "From sequence to function: the impact of the genome sequence on Drosophila biology." Briefings in Functional Genomics **11**(5): 333-335.

Saeys, Y., I. Inza, et al. (2007). "A review of feature selection techniques in bioinformatics." Bioinformatics **23**(19): 2507-2517.

Saha, I., U. Maulik, et al. (2011). "Fuzzy clustering of physicochemical and biochemical properties of amino Acids." Amino Acids: 1-12.

Sahu, S. S., G. Panda, et al. (2009). Improved protein structural class prediction using genetic algorithm and artificial immune system. 2009 World Congress on Nature and Biologically Inspired Computing, NABIC 2009 - Proceedings.

Sakar, O., O. Kursun, et al. (2010). Prediction of protein sub-nuclear location by clustering mRMR ensemble feature selection. Proceedings - International Conference on Pattern Recognition.

Sarda, D., G. Chua, et al. (2005). "pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties." BMC Bioinformatics **6**(1): 152.

Seker, H. (2008). "Novel weighted amino acid composition for prediction of protein structural classes within the context of multi-sensor data fusion approach." 8th IEEE International Conference on Bioinformatics and Bioengineering: 589-594.

Shen, H.-B. and K.-C. Chou (2009). "Predicting protein fold pattern with functional domain and sequential evolution information." Journal of Theoretical Biology **256**(3): 441-446.

Shen, H. B. and K. C. Chou (2008). "PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition." Analytical Biochemistry **373**(2): 386-388.

Smialowski, P., D. Frishman, et al. "Pitfalls of supervised feature selection." Bioinformatics **26**(3): 440-443.

Sneath, P. H. A. (1966). "Relations between chemical structure and biological activity in peptides." Journal of Theoretical Biology **12**(2): 157-195.

Tanford, C. (1962). "Contribution of hydrophobic interactions to the stability of the globular conformation of proteins." Journal of the American Chemical Society **84**(22): 4240-4247.

Tomii, K. and M. Kanehisa (1996). "Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins." Protein Engineering **9**(1): 27-36.

Wang, F., Z. Wang, et al. (2011). "Prediction of protein structural classes using the theory of increment of diversity and support vector machine." Wuhan University Journal of Natural Sciences **16**(3): 260-264.

Wang, X. (2003). "Feature Extraction and Dimensionality Reduction in Pattern Recognition and Their Application in Speech Recognition." Griffith University.

Wang, Z. X. and Z. Yuan (2000). "How good is prediction of protein structural class by the component- coupled method?" Proteins: Structure, Function and Genetics **38**(2): 165-175.

Wei-Shu Bu, Z.-P. F. Z. Z. C.-T. Z. (1999). "Prediction of protein (domain) structural classes based on amino-acid index." European Journal of Biochemistry **266**(3): 1043-1049.

Whitfield, E. J., M. Pruess, et al. (2006). "Bioinformatics database infrastructure for biotechnology research." Journal of Biotechnology **124**(4): 629-639.

Wilkins, M. R., E. Gasteiger, et al. (1999). "Protein identification and analysis tools in the ExPASy server." Methods in molecular biology (Clifton, N.J.) **112**: 531-552.

Wu, X., V. Kumar, et al. (2008). "Top 10 algorithms in data mining." Knowledge and Information Systems **14**(1): 1-37.

Xia, X.-Y., M. Ge, et al. (2012). "Accurate Prediction of Protein Structural Class." Plos One **7**(6).

Xiao, X., S. H. Shao, et al. (2006). "Using pseudo amino acid composition to predict protein structural classes: Approached with complexity measure factor." Journal of Computational Chemistry **27**(4): 478-482.

Xu, Y., J. Liu, et al. (2008). F-score Feature Selection Method May Improve Texture-based Liver Segmentation Strategies.

Yang, J.-Y., Z.-L. Peng, et al. (2009). "Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation." Journal of Theoretical Biology **257**(4): 618-626.

Yang, J. Y., Z. L. Peng, et al. (2010). "Prediction of protein structural classes for low-homology sequences based on predicted secondary structure." BMC Bioinformatics **11**(SUPPLL.1).

Yousef, M., M. Ketany, et al. (2009). "Classification and biomarker identification using gene network modules and support vector machines." BMC Bioinformatics **10**.

Yu, L. and H. Liu (2004). "Efficient feature selection via analysis of relevance and redundancy." Journal of Machine Learning Research **5**: 1205-1224.

Yu, T., Z.-B. Sun, et al. (2007). "Structural class tendency of polypeptide: A new conception in predicting protein structural class." Physica A: Statistical Mechanics and its Applications **386**(1): 581-589.

Zhang, C. T. and K. C. Chou (1992). "An optimization approach to predicting protein structural class from amino acid composition." Protein Science **1**(3): 401-408.

Zhang, T.-L., Y.-S. Ding, et al. (2008). "Prediction protein structural classes with pseudo-amino acid composition: Approximate entropy and hydrophobicity pattern." Journal of Theoretical Biology **250**(1): 186-193.

Zhang, T. L. and Y. S. Ding (2007). "Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes." Amino Acids **33**(4): 623-629.

Zhang, T. L., R. Wei, et al. (2007). "Using fuzzy support vector machine network to predict low homology protein structural classes." Pattern Recognition in Bioinformatics, Proceedings **4774**: 98-107.

Zhang, Y., C. Ding, et al. (2008). "Gene selection algorithm by combining reliefF and mRMR." BMC Genomics **9**(SUPPL. 2).

Zhang, Z. H., Z. H. Wang, et al. (2005). Nearest neighbor algorithm for prediction of protein domain structural class. Proceedings - Eighth International Conference on High-Performance Computing in Asia-Pacific Region, HPC Asia 2005, Beijing.

Zhao, G. and E. London (2006). "An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: Relationship to biological hydrophobicity." Protein Science **15**(8): 1987-2001.

Zhou, G. P. and N. Assa-Munt (2001). "Some insights into protein structural class prediction." Proteins-Structure Function and Genetics **44**(1): 57-59.

Zviling, M., H. Leonov, et al. (2005). "Genetic algorithm-based optimization of hydrophobicity tables." Bioinformatics **21**(11): 2651-2656.

Appendix I – Sequence Driven Features

Appendix I contains the feature names of 1497 sequence driven feature sued in chapter 4. Feature name in bold indicates a feature group.

Feature Index Number	Feature	Feature Size	Range
1	**Amino Acid Composition**	20	1-20
2	**Dipeptide Composition**	400	21-420
3	**Normalized Moreau-Borto Autocorrelation**	240	421-660
3.1	Hydrophobicity scale	30	*421-450*
3.2	Flexibility Indices	30	*451-480*
3.3	Polarizibility	30	*481-510*
3.4	Free energy in water	30	*511-540*
3.5	Residue accessibility surface area in Tripeptide	30	*541-570*
3.6	Residue volume	30	*571-600*
3.7	Steric parameter	30	*601-630*
3.8	Relative mutability	30	*631-660*
4	**Moran Autocorrelation**	240	661-900
4.1	Hydrophobicity scale	30	*661-690*
4.2	Flexibility Indices	30	*691-720*
4.3	Polarizibility	30	*721-750*
4.4	Free energy in water	30	*751-780*
4.5	Residue accessibility surface area in Tripeptide	30	*781-810*
4.6	Residue volume	30	*811-840*
4.7	Steric parameter	30	*841-870*
4.8	Relative mutability	30	*871-900*
5	**Geary autocorrelation**	240	901-1140
5.1	Hydrophobicity scale	30	*901-930*
5.2	Flexibility Indices	30	*931-960*
5.3	Polarizibility	30	*961-990*
5.4	Free energy in water	30	*991-1020*
5.5	Residue accessibility surface area in Tripeptide	30	*1021-1050*
5.6	Residue volume	30	*1051-1080*
5.7	Steric parameter	30	*1081-1110*
5.8	Relative mutability	30	*1111-1140*
6	**Composition, Transition & Distribution**	147	1141-1287
6.1	**Composition**	21	1141-1161
6.1.1	Hydrophobicity	3	*1141-1143*
6.1.2	Normalized van der waal volume	3	*1144-1146*
6.1.3	Polarity	3	*1147-1149*
6.1.4	Polarizability	3	*1150-1152*
6.1.5	Charge	3	*1153-1155*
6.1.6	Secondary structure	3	*1156-1158*
6.1.7	Solvent accessibility	3	*1159-1161*
7.1	**Transition**	21	1162-1182
7.1.1	Hydrophobicity	3	*1162-1164*
7.1.2	Normalized van der waal volume	3	*1165-1167*
7.1.3	Polarity	3	*1168-1170*
7.1.4	Polarizability	3	*1171-1173*
7.1.5	Charge	3	*1174-1176*
7.1.6	Secondary structure	3	*1177-1179*
7.1.7	Solvent accessibility	3	*1180-1182*
8.1	**Distribution**	105	1183-1287
8.1.1	Hydrophobicity	15	*1183-1197*
8.1.2	Normalized van der waal volume	15	*1198-1212*
8.1.3	Polarity	15	*1213-1227*
8.1.4	Polarizability	15	*1228-1242*
8.1.5	Charge	15	*1243-1257*
8.1.6	Secondary structure	15	*1258-1272*
8.1.7	Solvent accessibility	15	*1273-1287*
9	**Sequence Order**	160	1288-1447
9.1	**Sequence-order-coupling number**	60	1288-1347

9.1.1	Based on Schneider -Wrede distance	30	*1288-1317*
9.1.2	Based on normalized Granthan chemical distance	30	*1318-1347*
9.2	**Quasi-sequence-order descriptors**	100	1348-1447
9.2.1	Based on Schneider -Wrede distance	50	*1348-1397*
9.2.2	Based on normalized Granthan chemical distance	50	*1398-1447*
10	**Pseudo amino acid composition**	50	1448-1497
10.1	Weighted AAC	20	*1448-1467*
10.2	Lamda	30	*1468-1497*
11	All features	1479	1-1497

Appendix II Full list of amino acid indices from the AAindex database

Amino acid indices from the AAindex1 database (Kawashima and Kanehisa 2000).

Amino Acid Index (AAI)	Index Name
1	ANDN920101
2	ARGP820101
3	ARGP820102
4	ARGP820103
5	BEGF750101
6	BEGF750102
7	BEGF750103
8	BHAR880101
9	BIGC670101
10	BIOV880101
11	BIOV880102
12	BROC820101
13	BROC820102
14	BULH740101
15	BULH740102
16	BUNA790101
17	BUNA790102
18	BUNA790103
19	BURA740101
20	BURA740102
21	CHAM810101
22	CHAM820101
23	CHAM820102
24	CHAM830101
25	CHAM830102
26	CHAM830103
27	CHAM830104
28	CHAM830105
29	CHAM830106
30	CHAM830107
31	CHAM830108
32	CHOC750101
33	CHOC760101
34	CHOC760102
35	CHOC760103
36	CHOC760104
37	CHOP780101
38	CHOP780201
39	CHOP780202
40	CHOP780203
41	CHOP780204
42	CHOP780205
43	CHOP780206
44	CHOP780207
45	CHOP780208
46	CHOP780209
47	CHOP780210
48	CHOP780211
49	CHOP780212
50	CHOP780213
51	CHOP780214
52	CHOP780215
53	CHOP780216
54	CIDH920101
55	CIDH920102
56	CIDH920103
57	CIDH920104
58	CIDH920105
59	COHE430101
60	CRAJ730101
61	CRAJ730102
62	CRAJ730103
63	DAWD720101
64	DAYM780101
65	DAYM780201
66	DESM900101
67	DESM900102
68	EISD840101
69	EISD860101
70	EISD860102
71	EISD860103
72	FASG760101
73	FASG760102
74	FASG760103
75	FASG760104
76	FASG760105

77	FAUJ830101
78	FAUJ880101
79	FAUJ880102
80	FAUJ880103
81	FAUJ880104
82	FAUJ880105
83	FAUJ880106
84	FAUJ880107
85	FAUJ880108
86	FAUJ880109
87	FAUJ880110
88	FAUJ880111
89	FAUJ880112
90	FAUJ880113
91	FINA770101
92	FINA910101
93	FINA910102
94	FINA910103
95	FINA910104
96	GARJ730101
97	GEIM800101
98	GEIM800102
99	GEIM800103
100	GEIM800104
101	GEIM800105
102	GEIM800106
103	GEIM800107
104	GEIM800108
105	GEIM800109
106	GEIM800110
107	GEIM800111
108	GOLD730101
109	GOLD730102
110	GRAR740101
111	GRAR740102
112	GRAR740103
113	GUYH850101
114	HOPA770101
115	HOPT810101
116	HUTJ700101
117	HUTJ700102
118	HUTJ700103
119	ISOY800101
120	ISOY800102
121	ISOY800103
122	ISOY800104
123	ISOY800105
124	ISOY800106
125	ISOY800107
126	ISOY800108
127	JANJ780101
128	JANJ780102
129	JANJ780103
130	JANJ790101
131	JANJ790102
132	JOND750101
133	JOND750102
134	JOND920101
135	JOND920102
136	JUKT750101
137	JUNJ780101
138	KANM800101
139	KANM800102
140	KANM800103
141	KANM800104
142	KARP850101
143	KARP850102
144	KARP850103
145	KHAG800101
146	KLEP840101
147	KRIW710101
148	KRIW790101
149	KRIW790102
150	KRIW790103
151	KYTJ820101
152	LAWE840101
153	LEVM760101
154	LEVM760102
155	LEVM760103
156	LEVM760104
157	LEVM760105
158	LEVM760106
159	LEVM760107
160	LEVM780101
161	LEVM780103
162	LEVM780104
163	LEVM780105
164	LEVM780106

165	LEWP710101
166	LIFS790101
167	LIFS790102
168	LIFS790103
169	MANP780101
170	MAXF760101
171	MAXF760102
172	MAXF760103
173	MAXF760104
174	MAXF760105
175	MAXF760106
176	MCMT640101
177	MEEJ800101
178	MEEJ800102
179	MEEJ810101
180	MEEJ810102
181	MEIH800101
182	MEIH800102
183	MEIH800103
184	MIYS850101
185	NAGK730101
186	NAGK730102
187	NAGK730103
188	NAKH900101
189	NAKH900102
190	NAKH900103
191	NAKH900104
192	NAKH900105
193	NAKH900106
194	NAKH900107
195	NAKH900108
196	NAKH900109
197	NAKH900110
198	NAKH900111
199	NAKH900112
200	NAKH900113
201	NAKH920101
202	NAKH920102
203	NAKH920103
204	NAKH920104
205	NAKH920105
206	NAKH920106
207	NAKH920107
208	NAKH920108
209	NISK800101
210	NISK860101
211	NOZY710101
212	OOBM770101
213	OOBM770102
214	OOBM770103
215	OOBM770104
216	OOBM770105
217	OOBM850101
218	OOBM850102
219	OOBM850103
220	OOBM850104
221	OOBM850105
222	PALJ810101
223	PALJ810102
224	PALJ810103
225	PALJ810104
226	PALJ810105
227	PALJ810106
228	PALJ810107
229	PALJ810108
230	PALJ810109
231	PALJ810110
232	PALJ810111
233	PALJ810112
234	PALJ810113
235	PALJ810114
236	PALJ810115
237	PALJ810116
238	PARJ860101
239	PLIV810101
240	PONP800101
241	PONP800102
242	PONP800103
243	PONP800104
244	PONP800105
245	PONP800106
246	PONP800107
247	PONP800108
248	PRAM820101
249	PRAM820102
250	PRAM820103
251	PRAM900101
252	PRAM900103

253 PRAM900104
254 PTIO830101
255 PTIO830102
256 QIAN880101
257 QIAN880102
258 QIAN880103
259 QIAN880104
260 QIAN880105
261 QIAN880106
262 QIAN880107
263 QIAN880108
264 QIAN880109
265 QIAN880110
266 QIAN880111
267 QIAN880112
268 QIAN880113
269 QIAN880114
270 QIAN880115
271 QIAN880116
272 QIAN880117
273 QIAN880118
274 QIAN880119
275 QIAN880120
276 QIAN880121
277 QIAN880122
278 QIAN880123
279 QIAN880124
280 QIAN880125
281 QIAN880126
282 QIAN880127
283 QIAN880128
284 QIAN880129
285 QIAN880130
286 QIAN880131
287 QIAN880132
288 QIAN880133
289 QIAN880134
290 QIAN880135
291 QIAN880136
292 QIAN880137
293 QIAN880138
294 QIAN880139
295 RACS770101
296 RACS770102
297 RACS770103
298 RACS820101
299 RACS820102
300 RACS820103
301 RACS820104
302 RACS820105
303 RACS820106
304 RACS820107
305 RACS820108
306 RACS820109
307 RACS820110
308 RACS820111
309 RACS820112
310 RACS820113
311 RACS820114
312 RADA880101
313 RADA880102
314 RADA880103
315 RADA880104
316 RADA880105
317 RADA880106
318 RADA880107
319 RADA880108
320 RICJ880101
321 RICJ880103
322 RICJ880104
323 RICJ880105
324 RICJ880106
325 RICJ880107
326 RICJ880108
327 RICJ880109
328 RICJ880110
329 RICJ880111
330 RICJ880112
331 RICJ880113
332 RICJ880114
333 RICJ880115
334 RICJ880116
335 RICJ880117
336 ROBB760101
337 ROBB760102
338 ROBB760103
339 ROBB760104
340 ROBB760105

341	ROBB760106
342	ROBB760107
343	ROBB760108
344	ROBB760109
345	ROBB760110
346	ROBB760111
347	ROBB760112
348	ROBB760113
349	ROBB790101
350	ROSG850101
351	ROSG850102
352	ROSM880101
353	ROSM880102
354	ROSM880103
355	SIMZ760101
356	SNEP660101
357	SNEP660102
358	SNEP660103
359	SNEP660104
360	SUEM840101
361	SUEM840102
362	SWER830101
363	TANS770101
364	TANS770102
365	TANS770103
366	TANS770104
367	TANS770105
368	TANS770106
369	TANS770107
370	TANS770108
371	TANS770109
372	TANS770110
373	VASM830101
374	VASM830102
375	VASM830103
376	VELV850101
377	VENT840101
378	VHEG790101
379	WARP780101
380	WEBA780101
381	WERD780101
382	WERD780102
383	WERD780103
384	WERD780104
385	WOEC730101
386	WOLR810101
387	WOLS870101
388	WOLS870102
389	WOLS870103
390	YUTK870101
391	YUTK870102
392	YUTK870103
393	YUTK870104
394	ZASB820101
395	ZIMJ680101
396	ZIMJ680102
397	ZIMJ680103
398	ZIMJ680104
399	ZIMJ680105
400	AURR980101
401	AURR980102
402	AURR980103
403	AURR980104
404	AURR980105
405	AURR980106
406	AURR980107
407	AURR980108
408	AURR980109
409	AURR980110
410	AURR980111
411	AURR980112
412	AURR980113
413	AURR980114
414	AURR980115
415	AURR980116
416	AURR980117
417	AURR980118
418	AURR980119
419	AURR980120
420	ONEK900101
421	ONEK900102
422	VINM940101
423	VINM940102
424	VINM940103
425	VINM940104
426	MUNV940101
427	MUNV940102
428	MUNV940103

429	MUNV940104
430	MUNV940105
431	WIMW960101
432	KIMC930101
433	MONM990101
434	BLAM930101
435	PARS000101
436	PARS000102
437	KUMS000101
438	KUMS000102
439	KUMS000103
440	KUMS000104
441	TAKK010101
442	FODM020101
443	NADH010101
444	NADH010102
445	NADH010103
446	NADH010104
447	NADH010105
448	NADH010106
449	NADH010107
450	MONM990201
451	KOEP990101
452	KOEP990102
453	CEDJ970101
454	CEDJ970102
455	CEDJ970103
456	CEDJ970104
457	CEDJ970105
458	FUKS010101
459	FUKS010102
460	FUKS010103
461	FUKS010104
462	FUKS010105
463	FUKS010106
464	FUKS010107
465	FUKS010108
466	FUKS010109
467	FUKS010110
468	FUKS010111
469	FUKS010112
470	MITS020101
471	TSAJ990101
472	TSAJ990102
473	COSI940101
474	PONP930101
475	WILM950101
476	WILM950102
477	WILM950103
478	WILM950104
479	KUHL950101
480	GUOD860101
481	JURD980101
482	BASU050101
483	BASU050102
484	BASU050103
485	SUYM030101
486	PUNT030101
487	PUNT030102
488	GEOR030101
489	GEOR030102
490	GEOR030103
491	GEOR030104
492	GEOR030105
493	GEOR030106
494	GEOR030107
495	GEOR030108
496	GEOR030109
497	ZHOH040101
498	ZHOH040102
499	ZHOH040103
500	BAEK050101
501	HARY940101
502	PONJ960101
503	DIGM050101
504	WOLR790101
505	OLSK800101
506	KIDA850101
507	GUYH850102
508	GUYH850104
509	GUYH850105
510	JACR890101
511	COWR900101
512	BLAS910101
513	CASG920101
514	CORJ870101
515	CORJ870102
516	CORJ870103

517	CORJ870104	
518	CORJ870105	
519	CORJ870106	
520	CORJ870107	
521	CORJ870108	
522	MIYS990101	
523	MIYS990102	
524	MIYS990103	
525	MIYS990104	
526	MIYS990105	
527	ENGD860101	
528	FASG890101	

Appendix III Full list of amino acid indices found through literature searches

529	Factor1NEW	Atchley, Zhao et al. 2005
530	Factor2NEW	Atchley, Zhao et al. 2005
531	Factor3NEW	Atchley, Zhao et al. 2005
532	Factor4NEW	Atchley, Zhao et al. 2005
533	Factor5NEW	Atchley, Zhao et al. 2005
534	HYDNEW	Chou 2008
535	MassNEW	Chou 2008
536	pK1NEW(a-CO2H)	Chou 2008
537	pK2NEW(NH3)	Chou 2008
538	pINEW(NH3)	Chou 2008
539	GES	Zviling et al. 2005
540	GASet1	Zviling et al. 2005
541	Set2	Zviling et al. 2005
542	Set3	Zviling et al. 2005
543	Rk	Huang et al. 2007
544	Rc	Huang et al. 2007
545	Ro	Huang et al. 2007
546	Rb	Huang et al. 2007
547	K0	Fernandez et al. 2007
548	Hp	Fernandez et al. 2007
549	P	Fernandez et al. 2007
550	pHi	Fernandez et al. 2007
551	pK'	Fernandez et al. 2007
552	Mw	Fernandez et al. 2007
553	Bl	Fernandez et al. 2007
554	Rf	Fernandez et al. 2007
555	Mu	Fernandez et al. 2007
556	Hnc	Fernandez et al. 2007
557	Esm	Fernandez et al. 2007
558	El	Fernandez et al. 2007
559	Et	Fernandez et al. 2007
560	Pc	Fernandez et al. 2007
561	Ca	Fernandez et al. 2007
562	F	Fernandez et al. 2007
563	Br	Fernandez et al. 2007
564	aN	Fernandez et al. 2007
565	aC	Fernandez et al. 2007
566	aM	Fernandez et al. 2007
567	V0	Fernandez et al. 2007
568	Nm	Fernandez et al. 2007
569	Nl	Fernandez et al. 2007
570	Hgm	Fernandez et al. 2007
571	ASAD	Fernandez et al. 2007
572	ASAN	Fernandez et al. 2007
573	dASA	Fernandez et al. 2007
574	dGh	Fernandez et al. 2007
575	GhD	Fernandez et al. 2007
576	GhN	Fernandez et al. 2007
577	dHh	Fernandez et al. 2007
578	-TdSh	Fernandez et al. 2007
579	dCph	Fernandez et al. 2007
580	dGc	Fernandez et al. 2007
581	dHc	Fernandez et al. 2007
582	-TdSc	Fernandez et al. 2007
583	dG	Fernandez et al. 2007
584	dH	Fernandez et al. 2007
585	-TdS	Fernandez et al. 2007
586	v	Fernandez et al. 2007
587	s	Fernandez et al. 2007
588	f	Fernandez et al. 2007
589	Pf-s	Fernandez et al. 2007
590	Fauchere-Pliska	Kurgan et al. 2007
591	Proscale_4	Wilkins et al. 1999
592	Proscale_7	Wilkins et al. 1999
593	Proscale_10	Wilkins et al. 1999
594	Proscale_11	Wilkins et al. 1999
595	Proscale_13	Wilkins et al. 1999
596	Proscale_15	Wilkins et al. 1999

597	Proscale_18	Wilkins et al. 1999
598	Proscale_21	Wilkins et al. 1999
599	Proscale_23	Wilkins et al. 1999
600	Proscale_24	Wilkins et al. 1999
601	Proscale_28	Wilkins et al. 1999
602	Proscale_30	Wilkins et al. 1999
603	Proscale_39	Wilkins et al. 1999
604	Proscale_41	Wilkins et al. 1999
605	Proscale_42	Wilkins et al. 1999
606	Proscale_45	Wilkins et al. 1999
607	Proscale_47	Wilkins et al. 1999
608	Proscale_52	Wilkins et al. 1999
609	Proscale_53	Wilkins et al. 1999
610	Proscale_57	Wilkins et al. 1999
611	allergen1	Asakawa et al. 2010

Appendix IV Generated amino acid indices using SINGLE Linkage and Minimum Cluster Distance = 1

Generated Index ID	Variance	Amino acid Indices clustered. Refer to Appendix II and III
1	1	1 17
2	1	3 4
3	1	5 564
4	1	8 598
5	1	12 13
6	1	15 59
7	1	21 79
8	1	22 80
9	1	39 225
10	1	42 566
11	1	52 346
12	1	65 135
13	1	85 110
14	1	93 421
15	1	98 228
16	1	100 230
17	1	105 234
18	1	106 428
19	1	115 153
20	1	118 588
21	1	124 175
22	1	139 608
23	1	154 157
24	1	166 275
25	1	176 555
26	1	196 455
27	1	198 208
28	1	205 465
29	1	212 529
30	1	233 252
31	1	236 347
32	1	251 527
33	1	269 270
34	1	271 281
35	1	282 361
36	1	292 293
37	1	315 556
38	1	335 364
39	1	340 341
40	1	362 515
41	1	376 473
42	1	390 391
43	1	392 393
44	1	394 431
45	1	397 549
46	1	402 417
47	1	407 411
48	1	409 414
49	1	429 430
50	1	439 440
51	1	447 448
52	1	460 607
53	1	462 463
54	1	482 484
55	1	483 499
56	1	501 502
57	1	513 570
58	1	531 533
59	1	582 585
60	1	583 584
61	0.9999	78 396 553
62	0.9989	169 474 548
63	0.9959	72 535 552 586
64	0.9953	191 193 195
65	0.903	516 517 518 519 520
66	0.8736	76 133 536
67	0.8708	179 180 480
68	0.8699	160 162 222
69	0.8492	190 192 194 199
70	0.8422	426 427 609
71	0.8218	158 350 573
72	0.8203	35 151 481 505

73	0.8203	101 163 224
74	0.804	215 216 314 380
75	0.7995	41 337 405 565
76	0.7953	16 420 434
77	0.7935	278 279 280
78	0.7903	66 67 603
79	0.7857	156 249 432
80	0.7225	488 489 490 491 493 494
81	0.6962	89 324 406
82	0.6873	55 56 57 58
83	0.672	50 122 366
84	0.6704	128 131 444 445 446 559 563
85	0.6572	120 171 365
86	0.6563	201 202 206 457
87	0.6532	68 312 386 504 510 534 539 574 575 576
88	0.6431	87 352 353 479 506
89	0.634	178 399 554 596
90	0.6303	140 408 413
91	0.6271	92 403 404 450
92	0.6144	34 127 129 508 572
93	0.6142	209 247 514 543
94	0.5417	2 108 132 355 441 579
95	0.5382	9 32 109 150 471 472 567
96	0.5182	240 241 242 546
97	0.4912	254 263 264 265 266 338
98	0.4638	51 123 367 589
99	0.4604	64 134 136 137 188 437 438 453 454 456 466 467 468 469 601 610
100	0.4483	69 77 239 512 590 594 605
101	0.43	7 62 343 345 348
	94	
102	0.4304	88 94 146 398 538 550
103	0.3959	38 119 138 170 223 336 599
104	0.3473	10 11 184 210 319 351 381 593
105	0.3358	24 37 40 47 53 104 107 121 161 164 226 227 253 287 288 289 290 372 560 600
106	0.3032	126 173 174 303 306 333 369 371 416
107	0.2699	14 111 113 143 147 148 149 181 182 214 238 295 296 385 387 422 423 424 435 458 459 461 486 487 507 521 522 523 524 525 526 528 544 562

Appendix V Generated amino acid indices using SINGLE Linkage and Minimum Cluster Distance = 0.65

Generated Index ID	Variance	Amino acid Indices clustered. Refer to Appendix II and III
1	1	1 17
2	1	2 132
3	1	3 4
4	1	5 564
5	1	8 598
6	1	9 109
7	1	10 319
8	1	12 13
9	1	15 59
10	1	21 79
11	1	22 80
12	1	24 560
13	1	35 505
14	1	37 227
15	1	38 223
16	1	39 225
17	1	40 53
18	1	41 565
19	1	42 566
20	1	50 366
21	1	51 123
22	1	52 346
23	1	56 58
24	1	64 137
25	1	65 135
26	1	67 603
27	1	68 534
28	1	72 552
29	1	76 133
30	1	77 590
31	1	85 110
32	1	89 324
33	1	92 450
34	1	93 421
35	1	98 228
36	1	100 230
37	1	104 107
38	1	105 234
39	1	106 428
40	1	111 385
41	1	115 153
42	1	118 588
43	1	119 170
44	1	120 171
45	1	124 175
46	1	127 508
47	1	128 563
48	1	134 454
49	1	139 608
50	1	140 408
51	1	151 481
52	1	154 157
53	1	156 432
54	1	161 253
55	1	162 222
56	1	163 224
57	1	166 275
58	1	174 371
59	1	176 555
60	1	179 180
61	1	181 295
62	1	182 296
63	1	184 593
64	1	188 601
65	1	190 192
66	1	191 193
67	1	196 455
68	1	198 208
69	1	201 457
70	1	205 465
71	1	209 247
72	1	212 529
73	1	215 216
74	1	233 252
75	1	236 347
76	1	238 387
77	1	240 241

78	1	251 527
79	1	264 265
80	1	269 270
81	1	271 281
82	1	279 280
83	1	282 361
84	1	289 290
85	1	292 293
86	1	315 556
87	1	335 364
88	1	340 341
89	1	343 348
90	1	350 573
91	1	352 479
92	1	362 515
93	1	376 473
94	1	386 504
95	1	390 391
96	1	392 393
97	1	394 431
98	1	396 553
99	1	397 549
100	1	398 550
101	1	399 554
102	1	402 417
103	1	403 404
104	1	407 411
105	1	409 414
106	1	420 434
107	1	426 427
108	1	429 430
109	1	437 438
110	1	439 440
111	1	445 446
112	1	447 448
113	1	459 461
114	1	460 607
115	1	462 463
116	1	471 472
117	1	474 548
118	1	482 484
119	1	483 499
120	1	488 493
121	1	501 502
122	1	510 539
123	1	512 605
124	1	513 570
125	1	514 543
126	1	516 517
127	1	518 519
128	1	522 523
129	1	524 525
130	1	531 533
131	1	574 575
132	1	582 585
133	1	583 584
134	0.6497	126 173 333 416

Appendix VI Generated amino acid indices using COMPLETE Linkage and Minimum Cluster Distance = 1

Generated Index ID	Variance	Amino acid Indices clustered. Refer to Appendix II and III
1	1	19 363
2	1	20 283
3	1	22 80
4	1	31 284
5	1	34 509
6	1	35 505
7	1	39 225
8	1	44 344
9	1	49 321
10	1	60 185
11	1	63 117
12	1	70 86
13	1	74 90
14	1	87 425
15	1	100 230
16	1	106 428
17	1	111 385
18	1	114 397 549
19	1	115 153
20	1	124 175
21	1	130 545
22	1	134&454
23	1	139 608
24	1	143 148
25	1	147 221
26	1	149 297
27	1	151 481
28	1	152 395
29	1	156 432
30	1	167 246
31	1	174 371
32	1	187 530
33	1	189 532
34	1	196 455
35	1	198 208
36	1	205 465
37	1	209 247
38	1	212 529
39	1	218 380
40	1	219 478
41	1	220 298
42	1	233 252
43	1	239 480
44	1	249 250
45	1	251 378 527
46	1	257 299
47	1	259 327
48	1	261 262
49	1	271 281
50	1	272 401
51	1	279 280
52	1	282 361
53	1	294 537
54	1	310 451
55	1	335 364
56	1	349 558
57	1	369 589
58	1	402 417
59	1	407 411
60	1	423 435
61	1	424 507
62	1	441 579
63	1	462 463
64	1	464 597
65	1	475 511
66	1	482 484
67	1	486 487
68	1	489 494
69	1	501 502
70	1	503 602
71	1	514 543
72	1	518 519
73	1	557 577
74	1	561 571

75	1	582 585
76	0.9996	399 554 596
77	0.999	521 522 523
78	0.9989	169 474 548
79	0.9972	8 142 598
80	0.9959	72 535 552 586
81	0.9953	191 193 195
82	0.9643	386 504 576
83	0.935	77 590 592
84	0.8982	179 180 595
85	0.8737	485 583 584
86	0.8652	98 228 326
87	0.8502	316 574 575
88	0.8492	190 192 194 199
89	0.8471	447 448 500
90	0.8382	286 289 290
91	0.8203	101 163 224
92	0.8121	443 444 445 446
93	0.8096	32 471 472 567
94	0.8068	188 601 610
95	0.7929	15 59 217
96	0.7903	66 67 603
97	0.7831	93 421 496
98	0.7611	12 13 177
99	0.7588	215 216 314
100	0.7531	390 391 392 393
101	0.7513	9 109 112 150
102	0.7466	516 517 520
103	0.7421	14 238 387
104	0.7383	315 318 556
105	0.7277	38 138 223
106	0.7266	5 260 564
107	0.7201	184 210 381 593
108	0.7196	513 569 570
109	0.7061	64 136 137 453 468
110	0.7043	51 123 367
111	0.7026	181 295 562
112	0.6955	43 269 270
113	0.6843	1 17 368 376 473
114	0.6829	83 213 388
115	0.6796	357 374 547
116	0.6762	264 265 266

117	0.673	2 108 132 355
118	0.672	50 122 366
119	0.6703	540 541 542
120	0.6647	56 57 58
121	0.6615	48 236 347
122	0.6563	201 202 206 457
123	0.6517	383 531 533 536
124	0.6469	65 135 172 382
125	0.6461	116 328 329
126	0.6438	128 131 379 559 563
127	0.6412	99 229 258
128	0.637	45 61 186
129	0.6365	36 243 384
130	0.6195	16 339 420 434
131	0.6182	6 102 231
132	0.6101	436 458 459 461
133	0.6055	120 171 308 365
134	0.603	21 79 155 356
135	0.6027	488 490 491 493
136	0.6016	256 437 438 439 440 466 467
137	0.6006	29 33 154 157
138	0.5969	10 11 319 351
139	0.5948	322 342 354
140	0.5884	254 263 338
141	0.5758	27 267 268 330
142	0.5724	3 4 244 245
143	0.5702	178 211 313
144	0.5696	81 118 588
145	0.5656	28 470 580
146	0.5655	119 170 336 599
147	0.5643	23 309 449
148	0.56	203 204 207
149	0.5536	144 460 607 611
150	0.5509	158 325 350 498 573 578
151	0.549	352 353 479 506
152	0.5416	68 312 510 534 539
153	0.5359	26 78 82 84 396 553
154	0.5346	176 302 442 555 581
155	0.5262	113 182 296 528
156	0.5202	69 197 512 594 605 606
157	0.5182	240 241 242 546
158	0.5155	24 37 47 227 560

159	0.4868	409 410 414 495
160	0.4698	104 107 121 161 164 253
161	0.4684	46 232 375 377
162	0.4678	127 129 508 572
163	0.4611	40 53 287 288 372
164	0.4573	96 159 358 394 431
165	0.4567	126 173 303 306 333 416
166	0.4552	166 168 274 275 276
167	0.4511	54 55 362 483 497 499 515
168	0.4452	97 160 162 222 305 568
169	0.4394	7 62 343 345 348
170	0.4321	214 422 524 525 526 544
171	0.4316	105 226 234 237 320
172	0.4278	141 255 340 341
173	0.4187	91 140 360 408 412 413
174	0.4157	92 165 370 403 404 450
175	0.4122	75 400 418 419
176	0.3917	18 42 331 332 415 566 604
177	0.3799	73 76 133 200 551
178	0.3709	41 89 125 323 324 337 405 406 565
179	0.345	88 94 95 145 146 398 538 550
180	0.3401	291 292 293 307 311 334 426 427 429 430 609
181	0.2808	25 52 85 110 278 285 301 304 346 359 389 477 591

Appendix VII Generated amino acid indices using AVERAGE Linkage and Minimum Cluster Distance = 0.65 and 1.0 (both generated same set of results)

Generated Index ID	Variance	Amino acid Indices clustered. Refer to Appendix II and III
1	1	12 13
2	1	18 604
3	1	25 591
4	1	27 330
5	1	31 284
6	1	35 505
7	1	39 225
8	1	40 53
9	1	44 334
10	1	52 346
11	1	65 135
12	1	74 90
13	1	75 418
14	1	86 509
15	1	92 450
16	1	100 230
17	1	104 107
18	1	106 428
19	1	114 397 549
20	1	121 226
21	1	124 175
22	1	139 608
23	1	145 547
24	1	149 297
25	1	151 481
26	1	152 395
27	1	156 432
28	1	167 246
29	1	187 237
30	1	196 455
31	1	198 208
32	1	205 465
33	1	209 247
34	1	213 388
35	1	219 478
36	1	220 298
37	1	233 252
38	1	249 250
39	1	254 338
40	1	267 268
41	1	269 270
42	1	271 281
43	1	289 290
44	1	305 568
45	1	310 451
46	1	322 354
47	1	335 364
48	1	403 404
49	1	407 411

50	1	429 430
51	1	441 579
52	1	462 463
53	1	464 597
54	1	470 580
55	1	489 494
56	1	501 502
57	1	513 570
58	1	514 543
59	1	518 519
60	1	582 585
61	0.9989	169 474 548
62	0.9957	8 424 598
63	0.9953	191 193 195
64	0.9885	444 445 446
65	0.9694	9 109 150
66	0.9643	386 504 576
67	0.9129	176 555 581
68	0.8737	485 583 584
69	0.8736	315 316 556
70	0.8655	113 182 296
71	0.8492	190 192 194 199
72	0.8457	362 482 484 515
73	0.8243	272 531 533
74	0.8236	221 460 607
75	0.8203	101 163 224
76	0.8096	32 471 472 567
77	0.7995	41 337 405 565
78	0.7934	29 154 157
79	0.7929	15 59 217
80	0.7903	66 67 603
81	0.7831	93 421 496
82	0.7744	483 497 499
83	0.7531	390 391 392 393
84	0.7466	516 517 520
85	0.7421	14 238 387
86	0.7379	141 340 341
87	0.7342	85 110 309
88	0.7307	211 313 377
89	0.7266	5 260 564
90	0.7224	142 423 435
91	0.7026	181 295 562
92	0.6983	125 373 587
93	0.6924	263 264 265 266
94	0.6873	55 56 57 58
95	0.6843	1 17 368 376 473
96	0.6813	143 147 148
97	0.673	2 108 132 355
98	0.672	50 122 366
99	0.6703	540 541 542
100	0.666	292 293 419
101	0.6461	116 328 329
102	0.6432	161 164 253 287 288
103	0.637	45 61 186
104	0.634	178 399 554 596
105	0.6281	197 475 476
106	0.6255	19 60 185
107	0.6195	16 339 420 434

108	0.6182	6 102 231
109	0.6167	97 160 162 222
110	0.6055	120 171 308 365
111	0.6027	488 490 491 493
112	0.5956	91 360 412
113	0.5952	311 426 427 530 609
114	0.5898	111 115 153 385 487
115	0.585	48 235 236 347
116	0.5843	229 257 258
117	0.5765	285 301 304
118	0.5761	24 37 227 372 560
119	0.5729	128 131 559 563
120	0.5684	400 402 417
121	0.5637	70 81 117 118 588
122	0.5546	22 33 72 80 535 552 586
123	0.5535	105 234 320 401
124	0.5441	98 99 228 326
125	0.5416	68 312 510 534 539
126	0.5272	20 244 283
127	0.519	89 323 324 406
128	0.5182	240 241 242 546
129	0.5143	140 261 262 408 413
130	0.5108	436 458 459 461 611
131	0.5105	3 4 245 492
132	0.492	34 127 129 212 508 529 572
133	0.488	278 279 280 359
134	0.4868	409 410 414 495
135	0.4736	130 447 448 500 545
136	0.4694	158 350 498 573 578
137	0.4552	166 168 274 275 276
138	0.4394	7 62 343 345 348
139	0.4053	21 26 78 79 82 84 155 356 396 553
140	0.4045	69 77 179 180 239 480 511 512 590 592 594 595 605 606
141	0.3992	36 243 357 384 557 574 575 577
142	0.389	251 352 353 378 479 486 506 527
143	0.3872	42 331 332 415 566
144	0.381	46 232 302 375 442
145	0.3655	214 422 507 521 522 523 524 525 526 528 544
146	0.3636	96 159 325 358 394 431
147	0.3544	38 119 138 170 223 336 363 599
148	0.3462	88 94 95 146 398 538 550
149	0.3224	73 76 133 200 536 551
150	0.3056	144 215 216 218 314 374 380 503
151	0.3032	126 173 174 303 306 333 369 371 416
152	0.3026	30 49 51 123 172 321 367 382 589
153	0.2968	64 134 136 137 188 189 201 202 203 204 206 207 437 438 439 440 453 454 456 457 466 467 468 469 532 601 602 610
154	0.2921	10 11 183 184 210 319 351 381 558 593
155	0.2765	23 282 294 342 361 389 449 477 537

Appendix VIII chapter 4 individual structural class of proteins results

Results highlighted in bold are the highest accuracy for the respective protein structural class. Feature index # are listed in appendix I.

Individual structural class performance using testing dataset 25PDB with 10-fold test procedure

Feature Index #	All-α	All-β	α/β	α+β	Overall
1	53.66%	47.39%	**37.54%**	27.23%	**41.69%**
6.1	52.94%	45.12%	37.21%	29.94%	41.55%
6.1.6	**58.40%**	**53.03%**	15.73%	23.82%	39.03%
10.1	43.24%	38.30%	33.42%	35.16%	37.77%
10	46.41%	39.44%	34.04%	29.92%	37.65%
11	40.05%	44.23%	30.87%	29.01%	36.33%
2	43.47%	50.78%	34.37%	14.53%	35.87%
3	52.26%	27.00%	17.14%	**40.83%**	35.33%
6.1.1	43.44%	36.74%	28.81%	30.16%	35.14%
4.1	41.63%	40.58%	23.84%	29.05%	34.36%

Individual structural class performance using testing dataset 25PDB with leave-one-out test procedure

Feature Index #	All-α	All-β	α/β	α+β	Overall
1	53.17%	48.53%	**37.21%**	27.66%	**41.91%**
6.1	51.81%	43.08%	35.47%	33.79%	41.37%
6.1.6	**59.50%**	**54.20%**	15.99%	22.90%	39.45%
10.1	43.67%	37.41%	34.01%	35.37%	37.83%
10	54.30%	40.82%	30.23%	20.41%	36.81%
11	40.72%	43.99%	31.98%	29.48%	36.81%
3	56.56%	27.21%	15.70%	**38.55%**	35.61%
9.1.2	46.15%	34.24%	29.65%	29.93%	35.31%
2	38.69%	47.85%	38.37%	16.55%	35.19%
6.1.1	38.01%	35.83%	29.94%	35.37%	35.07%

Individual structural class performance using testing dataset 25PDB with independent-sets test procedure

Feature Index #	All-α	All-β	α/β	α+β	Overall
2	54.07%	**62.59%**	**65.99%**	**61.68%**	**60.79%**
3	**58.60%**	54.20%	56.98%	58.28%	57.01%
1	53.62%	57.60%	60.76%	49.89%	55.16%
4	49.32%	56.69%	58.14%	56.69%	55.04%
5	50.68%	55.33%	59.30%	55.78%	55.04%
4.1	52.26%	53.97%	59.59%	53.29%	54.50%
3.1	52.49%	52.15%	61.92%	52.83%	54.44%
5.1	50.23%	55.78%	62.21%	51.25%	54.44%
11	52.94%	53.06%	52.62%	49.43%	51.98%
3.7	57.24%	47.39%	49.71%	50.11%	51.20%

Individual structural class performance using testing dataset 1189 with 10-fold Individual test procedure

Feature Index #	All-α	All-β	α/β	α+β	Overall
2	23.02%	45.92%	**63.64%**	**22.50%**	**41.42%**
1	39.16%	51.05%	53.64%	13.33%	41.00%
6.1	41.78%	**51.67%**	49.09%	13.33%	40.38%
6.1.6	46.78%	44.58%	44.85%	20.42%	39.76%
10.1	21.95%	42.81%	58.79%	18.75%	38.06%
10	32.40%	46.31%	48.48%	13.33%	36.79%
11	26.20%	35.96%	54.85%	22.50%	36.69%
4	21.65%	44.17%	54.55%	16.25%	36.51%
6	23.42%	34.26%	54.55%	24.58%	36.04%
3	**48.04%**	33.23%	40.61%	19.17%	35.39%

Individual structural class performance using testing dataset 1189 with leave-one-out Individual test procedure

Feature Index #	All-α	All-β	α/β	α+β	Overall
2	26.01%	46.23%	**65.15%**	19.17%	**41.84%**
1	42.15%	**54.79%**	50.91%	12.92%	41.75%
6.1	39.01%	47.60%	51.82%	17.50%	40.46%
10.1	30.04%	44.18%	57.88%	17.50%	39.54%
6.1.6	48.43%	49.32%	40.91%	14.58%	38.89%
10	26.46%	39.73%	59.09%	20.42%	38.62%
4	17.94%	39.73%	56.97%	**26.25%**	37.51%
3	**52.02%**	35.62%	42.12%	16.25%	36.68%
11	26.91%	34.93%	53.64%	22.92%	36.31%
6	26.91%	34.25%	52.42%	19.58%	35.02%

Individual structural class performance using testing dataset 1189 with independent-sets test procedure

Feature Index #	All-α	All-β	α/β	α+β	Overall
2	56.50%	**67.47%**	**70.00%**	57.92%	**63.87%**
5	59.64%	58.56%	63.33%	**63.33%**	61.29%
4	**60.54%**	58.22%	62.12%	59.58%	60.18%
3	59.64%	56.16%	61.52%	61.67%	59.72%
1	59.64%	65.41%	62.42%	40.83%	57.88%
11	55.61%	55.14%	60.00%	51.25%	55.85%
4.1	53.81%	56.85%	56.06%	52.92%	55.12%
5.1	52.47%	57.19%	56.67%	48.33%	54.10%
3.1	53.36%	52.05%	56.67%	52.50%	53.82%
9	49.33%	55.82%	55.15%	51.67%	53.36%

Individual structural class performance using testing dataset Astral25 with 10-fold Individual test procedure

Feature Index #	All-α	All-β	α/β	α+β	Overall
6.1	**44.89%**	46.86%	28.96%	45.64%	**41.40%**
1	42.59%	45.15%	28.76%	42.11%	39.45%
10	37.38%	40.34%	29.04%	45.31%	38.13%
10.1	39.77%	40.74%	28.41%	42.39%	37.76%
2	29.07%	37.99%	27.01%	**52.10%**	37.15%
6.1.6	40.47%	**47.89%**	28.25%	32.22%	36.75%
11	32.70%	34.68%	30.15%	46.06%	36.25%
6	29.69%	33.58%	29.64%	44.29%	34.71%
3	44.89%	27.06%	**37.12%**	30.86%	34.60%
6.1.7	34.90%	30.51%	24.67%	47.16%	34.59%

Individual structural class performance using testing dataset Astral25 with leave-one-out Individual test procedure

Feature Index #	All-α	All-β	α/β	α+β	Overall
6.1	**46.82%**	**49.37%**	50.14%	22.84%	**42.08%**
1	45.05%	46.93%	50.61%	24.08%	41.56%
10.1	43.55%	43.63%	47.22%	24.00%	39.47%
10	38.87%	40.64%	51.15%	23.79%	38.79%
2	29.15%	37.03%	**58.89%**	26.40%	38.67%
11	36.31%	39.07%	48.10%	25.60%	37.47%
6.1.6	40.72%	46.38%	35.69%	25.09%	36.58%
6	32.60%	39.15%	45.93%	23.57%	35.55%
6.1.7	35.42%	33.25%	47.22%	23.64%	35.12%
3	45.14%	27.52%	30.87%	**37.56%**	34.89%

Individual structural class performance using testing dataset Astral25 with independent-sets test procedure

Feature Index #	All-α	All-β	α/β	α+β	Overall
2	47.65%	**53.22%**	**47.38%**	29.99%	**44.30%**
1	51.82%	47.96%	38.98%	37.51%	43.37%
6.1	53.21%	49.41%	39.28%	29.49%	42.00%
11	44.02%	46.05%	37.95%	34.75%	40.30%
3	**56.84%**	41.14%	27.49%	**37.68%**	39.40%
5.1	50.96%	48.68%	31.61%	30.33%	39.32%
4	42.20%	44.23%	33.24%	37.51%	38.82%
4.1	48.29%	51.04%	29.48%	30.74%	38.82%
3.1	50.64%	44.50%	28.67%	35.09%	38.62%
6	43.91%	46.59%	34.05%	31.66%	38.44%

Individual structural class performance using testing dataset Astral40 with 10-fold test procedure

Feature Index #	All-α	All-β	α/β	α+β	Overall
6.1	39.86%	45.93%	46.95%	28.86%	**40.53%**
10	31.96%	36.28%	50.55%	25.62%	36.77%
1	40.68%	43.41%	44.70%	28.65%	39.37%
10.1	36.61%	38.79%	48.33%	26.32%	37.86%
2	26.75%	36.28%	**52.39%**	27.30%	36.68%
11	28.59%	34.59%	51.27%	29.35%	36.85%
6.1.6	39.84%	**48.03%**	34.42%	26.59%	36.79%
6	24.92%	33.61%	47.20%	31.19%	35.16%
6.1.7	28.39%	30.30%	48.77%	26.11%	34.19%
3	**42.28%**	26.40%	32.30%	**36.81%**	34.07%

Individual structural class performance using testing dataset Astral40 with leave-one-out test procedure

Feature Index #	All-α	All-β	α/β	α+β	Overall
1	43.43%	47.11%	53.61%	25.57%	**42.63%**
6.1	42.32%	**48.67%**	51.43%	23.35%	41.57%
2	32.92%	41.90%	**62.86%**	21.08%	40.74%
10.1	41.91%	46.06%	52.35%	17.62%	39.62%
11	32.23%	40.34%	53.95%	28.59%	39.58%
10	40.66%	47.34%	38.45%	22.65%	36.94%
6	25.73%	36.17%	50.46%	32.43%	37.23%
3	**47.03%**	29.28%	35.93%	**38.11%**	37.14%
6.1.6	40.66%	47.34%	38.45%	22.65%	36.94%
4.1	32.92%	37.44%	45.91%	25.08%	35.76%

Individual structural class performance using testing dataset Astral40 with independent-sets test procedure

Feature Index #	All-α	All-β	α/β	α+β	Table 4-13 Overall Accuracy
1	37.02%	**54.52%**	55.73%	16.64%	**41.38%**
2	28.34%	39.96%	**67.36%**	19.75%	40.27%
6.1	36.72%	46.72%	53.35%	**22.11%**	40.16%
11	33.31%	39.16%	57.24%	21.13%	38.50%
6	30.86%	36.80%	56.20%	20.67%	37.01%
6.1.6	43.40%	43.56%	42.71%	19.63%	37.01%
10.1	29.82%	47.27%	51.43%	14.74%	36.40%
3	**54.82%**	34.57%	40.69%	17.85%	36.09%
5	26.34%	42.75%	48.73%	20.61%	35.35%
10	27.00%	42.19%	49.82%	19.29%	35.32%

Appendix IX chapter 5 GAAC method individual structural class of proteins results

Results highlighted in bold are the highest accuracy for the respective protein structural class. Amino index numbers (AAI #) are listed in appendix II.

Individual structural class performance using testing dataset 25PDB with 10-fold test procedure

AAI #	All-α	All-β	α / β	α + β	Table 5-11 Overall Accuracy
414	58.15%	56.91%	58.39%	21.75%	**48.21%**
456	52.97%	47.82%	**65.96%**	20.39%	45.67%
495	55.00%	52.36%	49.74%	**26.50%**	45.67%
343	52.98%	53.95%	61.28%	17.68%	45.61%
198	58.64%	46.94%	65.93%	14.74%	45.45%
160	55.89%	51.47%	56.35%	20.18%	45.38%
408	52.28%	55.98%	57.59%	18.14%	45.32%
610	52.51%	47.17%	65.76%	20.17%	45.27%
467	55.70%	50.12%	62.23%	16.56%	45.23%
568	**59.30%**	**58.04%**	53.96%	11.32%	45.19%

Individual structural class performance using testing dataset 25PDB with leave-one-out test procedure

AAI #	All-α	All-β	α / β	α + β	Table 5-11 Overall Accuracy
414	56.79%	57.37%	59.59%	**22.22%**	**48.38%**
495	58.14%	54.65%	54.07%	20.86%	46.52%
408	53.39%	59.41%	59.59%	15.19%	46.16%
198	**64.25%**	47.17%	65.12%	11.79%	46.04%
456	55.88%	50.57%	62.21%	18.14%	45.80%
437	51.13%	51.93%	62.79%	20.86%	45.74%
348	51.36%	56.69%	60.47%	17.46%	45.68%
230	62.44%	**58.50%**	47.09%	14.74%	45.62%
305	58.37%	51.70%	56.10%	18.37%	45.56%
466	52.49%	48.98%	**64.24%**	20.63%	45.56%

Individual structural class performance using testing dataset 25PDB with independent-sets test procedure

AAI #	All-α	All-β	α / β	α + β	Table 5-11 Overall Accuracy
31	**60.86%**	**68.03%**	75.00%	**58.96%**	**65.17%**
64	59.95%	65.31%	72.97%	56.92%	63.25%
134	59.05%	66.89%	72.38%	55.56%	62.95%
188	59.05%	66.44%	72.97%	55.56%	62.95%
466	60.18%	64.63%	**75.29%**	53.74%	62.77%
201	54.30%	66.44%	73.55%	58.73%	62.65%
456	59.50%	64.17%	75.00%	54.65%	62.65%
467	60.86%	64.17%	72.38%	55.33%	62.65%
602	55.20%	67.35%	74.71%	55.56%	62.53%
454	59.05%	67.35%	72.09%	53.51%	62.47%

Individual structural class performance using testing dataset 1189 with 10-fold test procedure

AAI #	All-α	All-β	α / β	α + β	Table 5-11 Overall Accuracy
437	42.53%	**57.25%**	78.48%	14.58%	**51.24%**
143	37.40%	56.15%	79.09%	13.33%	49.79%
414	49.75%	55.81%	69.09%	15.83%	49.76%
466	41.16%	50.38%	79.09%	15.83%	49.59%
467	45.25%	53.14%	72.12%	**17.08%**	49.29%
464	**51.16%**	49.37%	77.88%	7.08%	49.03%
411	44.29%	54.71%	74.24%	11.67%	49.01%
302	43.71%	45.87%	**81.82%**	12.50%	48.95%
75	37.18%	55.55%	72.42%	19.17%	48.84%
154	46.27%	53.85%	72.42%	12.50%	48.78%

Individual structural class performance using testing dataset 1189 with leave-one-out test procedure

AAI #	All-α	All-β	α / β	α + β	Table 5-11 Overall Accuracy
437	46.19%	55.48%	76.67%	13.33%	**50.69%**
467	47.09%	55.14%	77.27%	12.08%	50.69%
411	45.74%	55.82%	76.97%	10.00%	50.05%
466	41.26%	53.08%	79.70%	13.75%	50.05%
143	36.77%	56.51%	80.30%	12.08%	49.86%
414	50.67%	**56.85%**	69.09%	14.17%	49.86%
464	**48.88%**	47.95%	80.91%	10.00%	49.77%
463	48.88%	49.66%	71.82%	**19.17%**	49.49%
302	48.43%	49.32%	79.70%	8.75%	49.40%
31	35.43%	46.92%	**87.58%**	12.08%	49.22%

Individual structural class performance using testing dataset 1189 with independent-sets test amino

AAI #	All-α	All-β	α / β	α + β	Table 5-11 Overall Accuracy
31	**66.37%**	**68.49%**	78.48%	54.58%	**68.02**
28	65.02%	63.36%	**80.00%**	53.75%	66.64
437	62.78%	65.75%	77.88%	55.83%	66.64
201	63.68%	63.36%	78.18%	56.25%	66.36
137	65.47%	65.07%	76.06%	54.17%	66.08
196	65.92%	61.99%	76.36%	56.67%	65.99
302	65.02%	64.38%	77.27%	52.92%	65.90
64	63.68%	65.07%	75.76%	55.00%	65.81
453	63.68%	65.07%	76.97%	52.92%	65.71
457	61.43%	62.33%	76.36%	**58.33%**	65.53

Individual structural class performance using testing dataset Astral25 with 10-fold test procedure

AAI #	All-α	All-β	α / β	α + β	Table 5-11 Overall Accuracy
437	50.37%	46.64%	71.42%	21.86%	**47.90%**
568	52.67%	**55.76%**	65.84%	17.22%	47.82%
466	47.19%	46.10%	**73.98%**	21.70%	47.76%
346	49.67%	52.63%	65.25%	22.77%	47.71%
414	53.91%	52.71%	64.75%	19.70%	47.69%
64	48.69%	49.24%	69.29%	22.22%	47.66%
409	48.44%	52.78%	69.77%	18.32%	47.58%
230	**54.27%**	50.34%	63.13%	**22.81%**	47.56%
137	50.02%	50.10%	67.53%	21.43%	47.45%
464	50.46%	43.10%	72.83%	21.86%	47.45%

Individual structural class performance using testing dataset Astral25 with leave-one-out test procedure

AAI #	All-α	All-β	α / β	α + β	Table 5-11 Overall Accuracy
466	50.97%	48.19%	77.27%	**19.87%**	**49.51%**
437	50.71%	47.88%	**79.92%**	16.24%	49.17%
136	49.38%	50.31%	78.97%	15.81%	49.10%
64	50.97%	50.86%	76.80%	15.81%	48.96%
137	51.59%	50.00%	75.10%	17.77%	48.93%
532	48.50%	52.28%	77.20%	15.74%	48.87%
414	**55.12%**	52.91%	70.90%	16.03%	48.75%
568	52.92%	**56.45%**	70.28%	15.01%	48.70%
464	52.30%	45.68%	75.44%	19.72%	48.64%
346	50.62%	52.99%	73.00%	16.90%	48.62%

Individual structural class performance using testing dataset Astral25 with independent-sets

AAI #	All-α	All-β	α / β	α + β	Table 5-11 Overall Accuracy
414	**63.16%**	**57.52%**	64.39%	23.26%	**51.74%**
31	45.89%	40.11%	54.82%	**29.17%**	42.76%
573	53.80%	52.24%	70.72%	25.77%	51.08%
468	55.78%	53.65%	68.85%	24.15%	50.85%
136	61.71%	50.75%	68.49%	22.12%	50.72%
466	57.65%	46.17%	71.80%	25.69%	50.70%
581	59.83%	55.15%	62.45%	26.01%	50.64%
532	59.52%	50.40%	70.58%	21.31%	50.59%
54	55.25%	49.08%	**72.16%**	23.82%	50.53%
346	54.73%	53.74%	67.77%	24.23%	50.36%

Individual structural class performance using testing dataset Astral40 with 10-fold test procedure - amino index numbers are listed in Appendix II

AAI #	All-α	All-β	α / β	α + β	Table 5-11 Overall Accuracy
568	49.34%	53.85%	70.86%	19.24%	**48.84%**
414	**52.74%**	54.38%	65.69%	20.11%	48.38%
160	48.58%	51.25%	**73.73%**	17.19%	48.35%
412	48.03%	**57.85%**	68.68%	16.92%	48.31%
495	48.94%	49.44%	71.39%	**20.97%**	48.29%
230	50.73%	50.45%	69.12%	20.38%	48.09%
138	45.33%	54.92%	69.17%	20.32%	48.07%
97	47.09%	53.18%	70.77%	18.49%	47.99%
162	46.51%	52.53%	70.63%	19.03%	47.82%
305	49.97%	51.72%	68.10%	19.84%	47.80%

Individual structural class performance using testing dataset Astral40 with leave-one-out test procedure

AAI #	All-α	All-β	α / β	α + β	Table 5-11 Overall Accuracy
64	46.82%	48.73%	78.98%	22.22%	**50.23%**
437	52.07%	50.75%	79.37%	15.73%	50.22%
532	50.41%	54.98%	77.92%	14.70%	50.22%
134	50.35%	51.79%	78.60%	16.59%	50.12%
468	45.23%	53.18%	80.05%	17.57%	50.09%
136	48.96%	52.03%	78.64%	17.08%	50.04%
414	**54.22%**	**55.73%**	73.17%	15.51%	50.01%
31	47.86%	50.41%	**83.20%**	14.11%	49.96%
568	51.45%	55.44%	73.70%	17.08%	49.94%
456	47.99%	46.30%	78.74%	**22.65%**	49.92%

Individual structural class performance using testing dataset Astral40 with independent-sets test procedure

AAI #	All-α	All-β	α / β	α + β	Table 5-11 Overall Accuracy
568	55.50%	53.21%	71.46%	**16.51%**	**49.43%**
414	**58.50%**	55.48%	69.38%	13.50%	49.20%
138	52.49%	55.54%	69.63%	13.55%	48.08%
170	55.13%	51.38%	70.14%	14.35%	47.96%
119	54.69%	53.77%	71.51%	10.42%	47.83%
412	53.96%	**57.44%**	69.22%	10.02%	47.80%
599	56.60%	54.19%	70.24%	9.85%	47.80%
437	48.53%	54.01%	**75.32%**	10.48%	47.77%
98	56.23%	52.36%	70.75%	10.82%	47.68%
32	42.82%	44.34%	81.01%	16.57%	47.52%